Ground Truth Evaluation for Event-Based Silicon Retina Stereo Data

Jürgen Kogler
AIT Austrian Institute of Technology GmbH
Donau-City-Strasse 1, 1220 Vienna, Austria
juergen.kogler.fl@ait.ac.at

Florian Eibensteiner
Upper Austria University of Applied Sciences
Softwarepark 11, 4232-Hagenberg, Austria
florian.eibensteiner@fh-hagenberg.at

Martin Humenberger
AIT Austrian Institute of Technology GmbH
Donau-City-Strasse 1, 1220 Vienna, Austria
martin.humenberger@ait.ac.at

Margrit Gelautz
Vienna University of Technology
Favoritenstrasse 9-11, 1040 Vienna, Austria
margrit.gelautz@tuwien.ac.at

Josef Scharinger
Johannes Kepler University Linz
Altenberger Strasse 69, 4040 Linz, Austria
josef.scharinger@jku.at

Abstract

In this paper we present a new approach for the evaluation of event-based Silicon Retina *stereo matching results. The evaluation of stereo matching algorithm results is a necessary task for the development, comparison, and improvement of depth generating camera systems. In contrast to conventional frame-based cameras, the silicon retina sensors delivers asynchronous events instead of synchronous intensity or color images. The polarity of the events represents either an increase (on-event) or a decrease (off-event) of the brightness of the projected scene point. This is the reason why existing ground truth data and evaluation platforms are not suitable for testing silicon retina stereo camera systems. For the analysis of the introduced novel evaluation method, we use an area-based (sum of absolute difference) algorithm for the event-driven sensor system. A conventional video camera stereo vision system is used to produce reference data. The results show that the presented method offers new opportunities for the evaluation of stereo matching results computed from silicon retina stereo data.*

1. Introduction

Automation, especially the therefore required embedded systems, makes our daily life activities easier and will not be missed any more. The car we drive is assembled nearly complete autonomously, driver assistance systems improve safety in traffic, our food and other articles of daily use come from large factories where production is done by robotic systems, and even some transport systems, such as rail shuttles on airports, drive completely autonomous without human interaction (to date in protected and closed areas only). Also embedded systems for self-driving cars are a topic of remarkable interest. The leading car manufacturers and technology providers world wide work hard on tackling this challenge.

Another application is health care, e.g. fall detection in nursing or private homes, to extend the time of independent living for elderly or sick people. Most of the autonomous systems which can accomplish the mentioned applications need 3D data of the environment to operate precisely and reliably.

State-of-the-art ranging sensors and technologies comprise active sensors, such as laser range finders, laser scanners, time-of-flight (TOF) cameras, ultrasonic detectors, radar, light-section, and structured light as well as passive technologies, such as structure from motion, optical flow, and stereo vision. Active sensors are based on emission and reception of light or radiation, while passive sensors try to reconstruct their environment by acquisition of the scene only. For the evaluation of the applicability of a sensor technology, it is essential to test the depth accuracy of the sensor itself. While it is pretty well known under which circumstances active sensors work best, there is still a gap in research for passive technologies.

Stereo vision, e.g., is based on capturing the area to observe from at least two different angles by digital cameras

and estimating the correspondences between these two images. Each visible scene point is projected onto the camera's image planes and represented by a pixel in the image frame. The horizontal displacement between the left camera's projection and the right camera's projection, representing the same scene point, is called *disparity*. A pixel's disparity is inversely proportional to the scene point's depth. The geometry between the calibrated cameras is known, which enables the determination of the 3D point in camera coordinates with the origin in the center of one of the cameras. Of course, only scene points visible in both images can be processed this way. The result of stereo vision based 3D measurement is a depth map and a 3D point cloud of the observed scene, which can then be used for further processing.

The crucial part of stereo vision is solving the so-called correspondence problem, also called stereo matching. Ambiguity and uncertainty associated with visual processing make this a challenging task. In particular, for embedded systems with restricted resources, used within the mentioned real-time applications, the computational expensive nature of stereo vision algorithms is a very sophisticated issue. Especially textureless areas are a significant problem because it is not possible to assign corresponding pixels to each other if all pixels have the same color and neighborhood.

Stereo matching is a well-known research topic, so lots of different approaches (varying in, e.g., feature or area analysis, processing time, and processing platforms) to solve the correspondence problem exist. A good comparison of different approaches as well as more basic information about stereo vision can be found in the work of Brown *et al*. [1] as well as in Scharstein and Szeliski [14].

1.1. Event-based Stereo Vision

A different approach of stereo vision is using two *Silicon Retina* sensors instead of conventional digital frame-based camera chips. In 1988, Mead and Mahowald [10] developed an electronic silicon model which reproduces the basic steps of human visual processing. One year later, Mahowald and Mead [9] implemented the first bio-inspired sensor based on silicon. Further developments of the silicon retina sensors are described in the work of Lichtsteiner *et al*. [6, 7].

By contrast to conventional frame-based image sensors, which generate frames of intensity or color values representing the observed area, these kinds of event-based neuromorphic visual sensors only deliver events on intensity changes caused by the dynamic parts of a scene. Hence, all static regions of a scene, e.g., the background, are suppressed, and therefore the transmitted event stream contains no redundant data, but only information of moving structures.

An event encodes the pixel's location on the chip, the timestamp when the event occurred, and the polarity. The polarity of the produced events can be either 1 (on-event) or -1 (off-event), when a positive or a negative change of illumination was detected, respectively. Each pixel of the sensor measures the changes of illumination in a logarithmic way and works asynchronously and time-continuously.

The data format is called *Address-Event-Representation* (AER) protocol, which has been introduced by Sivilotti [16] and Mahowald [8] in order to model the transmission of neural information within biological systems. The data transmission is completely frame-free and asynchronous, which means that events are transmitted without a fixed rate.

These characteristics and the efficient event generation yield to a very fast vision sensor system with a high temporal resolution and a high dynamic range. Hence, this kind of sensor technology is very suitable for high-speed real-time applications in uncontrolled environments with varying lighting conditions implemented on embedded systems with limited resources.

In terms of stereo vision, the same rules and challenges apply for event-based systems as apply for conventional frame-based systems. Here, as well, the solving of the correspondence problem is the major issue.

For evaluation of stereo matching results, consequently the accuracy of the stereo vision system, exact reference data of the 3D geometry of the observed scene is needed. This reference data or reference depth map (also called *Ground Truth*) usually consists of dense disparity data, or depth information of the scene, respectively, which allows a pixel-based evaluation of stereo matching results. Unfortunately, event-based vision sensors make matters worse because a dynamic scene or moving structures are needed for event generation. However, this demands an additional stereoscopic survey of the dynamic scene to gather the 3D depth information which can be used as reference data.

Due to the lack of evaluation methods for event-driven stereo systems, we propose in this paper a novel approach for testing event-based stereo matching algorithms. To do this, we combine a silicon retina stereo sensor with a conventional frame-based stereo vision camera and use the stereo matching results of the conventional camera system as reference information for the event-based stereo sensor. In this way, we can create our own datasets and evaluate the characteristics of different sensor settings and their influence on stereo matching in detail.

The remainder of the paper is organized as follows. In section 2 the related work of testing stereo matching results with ground truth data is discussed. Section 3 gives an overview of the ground truth setup evaluated in this work. In section 4 we evaluate the introduced approach with real world data, and in the final section 5 a conclusion with an outlook of future work is given.

2. Related Work

As a matter of fact, evaluation data for silicon retina based stereo vision algorithms with pixel-based reference data is rarely available. Contrary, stereo vision with conventional frame-based cameras is a well investigated research topic, thus, also evaluation methods of stereo matching approaches came up in the last ten years. In the next section, representative methods will be briefly described, which build the basics of the new approach presented in this paper. We appreciate all the excellent work that has been done in this research field, but here, we will focus on selected approaches.

2.1. Stereo Matching Evaluation

The most popular stereo matching evaluation platform is the Middlebury stereo database [1]. Scharstein and Szeliski [14] developed an online evaluation platform which provides a large number of stereo image datasets consisting of the stereo image pair and the appropriate ground truth data. The datasets represent static scenes and are created with a structured light approach [15]. To evaluate an algorithm on this website, disparity maps of all four datasets have to be generated and uploaded. The evaluation engine calculates the percentage of badly matched pixels (false positives), within a certain error threshold, by pixel-wise comparison with the reference image. Many stereo algorithm developers, approximately 145 entries to date, have used this platform for evaluation. This gives a significant overview of how the developed algorithm performs in comparison to other algorithms. The platform is up-to-date and constantly growing.

The disadvantage of the Middlebury platform is that the datasets do not realistically represent real-world scenarios stereo matching algorithms have to deal with in real applications. Especially in driver assistance systems, for autonomous robotics as well as consumer vehicles, the usage of stereo vision as 3D sensor technology has been growing over the last couple of years. To provide a suitable evaluation platform for especially this kind of application, KITTI [2] was introduced by Geiger *et al*. [2].

Similar to Middlebury, this platform provides datasets to evaluate stereo vision algorithms (as well as optical flow, tracking, odometry, and object detection) online. Contrary to Middlebury, these datasets are recorded from the roof of a car, driving on regular roads, with a front pointing stereo camera. The reference 3D data is determined with a laser scanner calibrated onto the stereo camera. Another remarkable difference to the Middlebury database is that at KITTI the processing time of the algorithm is also a part of the evaluation. This platform is rather new in the stereo vision community, thus, fewer algorithms are available than at Middlebury.

The Auckland Image Sequence Analysis Test Site (EISATS)[3] provides several datasets of, e.g., dangerous situations in traffic. These datasets include also challenging scenes for the camera hardware, such as direct sunlight, shadows, and fluctuating light. Unfortunately, no reference data is available which makes no direct statistical evaluation and comparison of different algorithms possible. To overcome this limitation, EISATS also provides synthetic sequences of automotive scenes [18].

A further stereo vision evaluation method was presented by Meister *et al*. [11]. The provided datasets show a huge variety of different weather conditions, motion, and depth layers. City as well as countryside situations were acquired at night and at day [4]. The provided reference stereo data is determined with a semi-global matching approach and is a feasible method to evaluate stereo matching algorithms.

A first evaluation platform for event-driven stereo vision algorithms was introduced by Sulzbachner *et al*. [17]. The test suite generates both, synthetic event data streams for the stereo matching algorithms, as well as the ground truth data for the subsequent evaluation step. For the event data generation no exact model of the behavior of a silicon retina is used and therefore, the asynchronous characteristics are not considered and algorithms can not be tested under real world conditions.

All the presented evaluation methods and platforms well contributed to make progress in stereo vision for scientific as well as industrial purposes. However, none of them can be used for evaluating silicon retina stereo systems under real conditions because of the following reasons. First, many datasets are static, so no events can be created and, thus, no silicon retina output is available.

However, second (and this describes why it does not work for us) all datasets acquired with conventional video cameras do not represent the asynchronous, time-continuous, event-driven spiking output of the silicon retina chip properly. Challenging lighting conditions and fast moving objects can be well handled by a silicon retina sensor and are a less significant limitation than for conventional video cameras. That is why we had to develop an approach which uses real silicon retina output data for evaluation.

All these reasons motivated us to develop a new evaluation approach for event-based stereo vision algorithms. The usage of a highly sophisticated stereo matching algorithm for reference data generation is a promising approach and leads us towards our novel solution presented in this paper.

[1] http://vision.middlebury.edu/stereo/

[2] http://www.cvlibs.net/datasets/kitti/index.php

[3] http://www.mi.auckland.ac.nz/EISATS

[4] http://hci.iwr.uni-heidelberg.de/Benchmarks/document/Challenging_Data_for_Stereo_and_Optical_Flow/

3. Ground Truth System Setup

The ground truth generation for the event-driven stereo vision system is based on a conventional stereoscopic vision system with two grayscale cameras. For this reason the used grayscale camera system has to achieve at least twice of the accuracy of the retina system, in order to use the generated disparity information as a reference for the event-based systems under test. In the following figure 1, the stereo system in the white box represents the silicon retina stereo system, which is under test. This system has a rigid connection to the grayscale reference stereo vision system above, which is shown in the dashed bounding box.

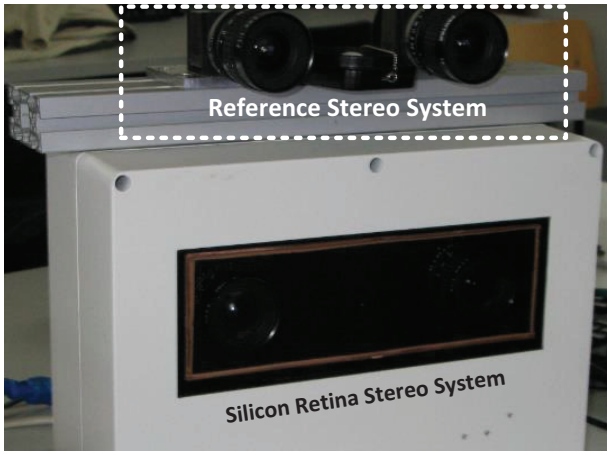


Figure 1. Camera setup for ground truth generation. The white box holds the embedded silicon retina stereo system which has a rigid connection to the stereo system in the dashed bounding box above, which acts as the reference stereo camera system.

3.1. Silicon Retina Stereo Sensor

As already mentioned above, a silicon retina sensor generates events if the intensity in the observed scene changes, whereby an event encodes the pixel's location, the time of origin of the detected change, and the polarity. The sensor system we used in this work features a spatial resolution of 304x240 pixels, a temporal resolution up to 10ns, and a dynamic range of 143dB. Further, and more detailed technical information can be found in the work of Posch *et al*. [12].

An event is given as [13]: $e(\mathbf{p}, t)$, where t is the time of occurrence and $\mathbf{p} = (x, y)^T$ the spatial location of the pixel which fires the event. However, an event can be set to the following values

$$e(\mathbf{p}, t) = \begin{cases} +1 & I(\mathbf{p}, t) > I(\mathbf{p}, t - \Delta t) \\ -1 & I(\mathbf{p}, t) < I(\mathbf{p}, t - \Delta t) \end{cases}, \qquad (1)$$

depending on whether a positive or negative change of illumination I over a period of time Δt was detected. Each pixel of the sensor measures the intensity in an asynchronous, time-continuous, and logarithmic way as following [5]

$$\frac{d}{dt} log I = \frac{\frac{dI}{dt}}{I}. \qquad (2)$$

If no change of intensity was detected, no events are generated and therefore no event data is transmitted. Thus, static parts of the scene, e.g., background information, are completely suppressed by the sensor.

Before the depth map used in the test section can be calculated by the stereo vision algorithm, the events transmitted from the silicon retina cameras have to be converted into grayscale images. Since events represent intensity changes over time, they can easily be transformed into an grayscale image by using the timing information for accumulating them to frames, and the polarity to gather a gray value for each pixel. Figure 2 shows the dataflow which generates the grayscale images used in this work.

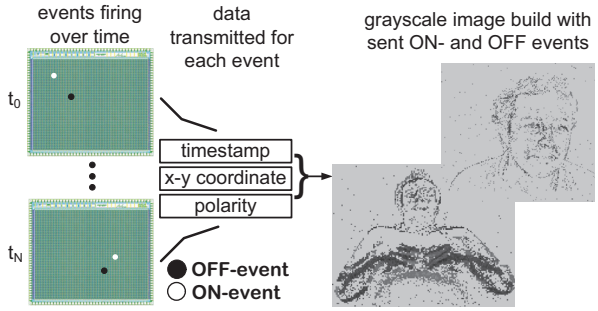


Figure 2. Grayscale image generation: Events send time, coordinates and polarity. Each event contributes to a grayscale value in the grayscale image which collects all events within a certain time (history).

For computing depth information from the generated grayscale images, we use an area-based stereo matching algorithm, which applies the *Sum of Absolute Differences* (SAD) as cost function. The time used for generating grayscale images is called history, and is, besides the window size of the matching kernel, the second configurable parameter for the cost aggregation. The output of the algorithm is a sparse disparity map, containing depth values achieved from the input data. A more detailed information of the used stereo matching approach can be found in the work of Kogler *et al*. [4].

3.2. Grayscale Stereo Sensor (reference stereo)

The reference system used in this paper is based on two *Imaging Development Systems* [5] (IDS) cameras (model UI-1220SE-M-GL Rev.2) which are mounted on a rigid baseline of 0.12m. The cameras transmit their images to the PC, where the further processing takes place. The rectified images are processed with an accurate and reliable

[5] http://en.ids-imaging.com/

distance	avg err	error
1.0m	0.012m	1.20%
1.5m	0.017m	1.13%
2.0m	0.027m	1.35%
2.5m	0.040m	1.60%
3.0m	0.081m	2.69%
3.5m	0.117m	3.35%
4.0m	0.220m	5.48%

Table 1. Evaluation of the distance accuracy calculated from the reference stereo vision system. In the testing range, the depth algorithm output is compared with the real distance measured by a laser distance meter.

census-based stereo matching algorithm. Details about the stereo matching engine can be found in the work of Humenberger *et al*. [3].

For the measurement of the reference system's accuracy, objects were placed at different distances. All distances were measured with a laser measurement device and compared with the depth output of the stereo algorithm. The accuracy was evaluated in the range where the tests took place. In the experimental results section, test data in the range of 1m to 4m was used, which is the range where the accuracy of the reference system was also measured. The average distance error in the range of interest is shown in table 1. The results show that the reference system has in close distances till 2.5m an error of less than 1.6%, and up to 4m less than 5.48%.

3.3. Configuration of the Stereo Heads

For our purpose both stereo camera systems need to be calibrated and registered onto each other, in a way that they have a pixel congruent representation of the scene in front of them.

3.3.1 Calibration

Before the stereo vision systems are ready to use, the stereo heads are calibrated separately to each other. The reference system as well as the silicon retina stereo system use the same calibration procedure as described in the work of Zhang [19]. The only difference is the pattern used. For the reference system the classic checkerboard calibration pattern is captured in different positions to provide necessary feature points. In contrast, the silicon retina system uses a circle pattern flashing on a computer display to generate stimuli for the retina sensors and later to extract the feature points for the calibration step. In this case the computer display or the silicon retina stereo camera can be moved to capture the necessary different views. After the calibration step, for both stereo heads all calibration and rectification parameters are available, which are further used in the registration process.

3.3.2 Registration

For the registration of both cameras onto each other and the achievement of a common understanding of the scene, the left image is used as reference for the depth map. Therefore, the registration will take place for the left view of the stereo systems.

In a first setup we tried to register both camera systems to a common world coordinate system. The results have shown that the accuracy depends on the exact calculation of the origin point and orientation of the world coordinate system. This lack of accuracy leaded us to another approach with more promising results. The approach introduced in this work is based on homographs, which represent the projective transformation between two planar spaces. Due to this fact, we use a homography H

$$p_{ref} = H \cdot p_{sr} \tag{3}$$

with $p_{sr}, p_{ref} \in \mathbb{R}^2$ to connect both camera systems to each other, and transform a point p_{sr} from the silicon retina stereo camera to a point p_{ref} in the reference stereo camera system. The homography H is determined to match a certain plane for the given feature points shown in figure 3. The

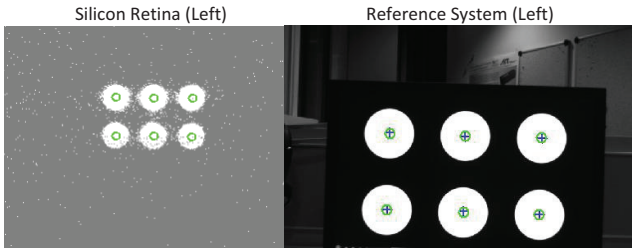


Figure 3. Left: silicon retina camera image, Right: reference image. - The circles represent the exact feature points extracted and crosses show the feature points transformed with the homography.

left image shows the silicon retina image with the feature points p_{sr} from the left sensor, and the right image shows the reference image feature points p_{ref} from the left camera. All feature points represented by the green circles are the exact extracted feature points. The blue crosses in the right image, are the transformed feature points, estimated with the calculated homography at a distance of 1m.

Using only this homography will lead to errors by applying it to other distances. This is the reason why the homography H was calculated at different distances d (in meter), where $d \in M := \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$. For each of these distances the corresponding homography $H(d)$ is calculated using the singular value decomposition (SVD) given by

$$[USV] = SVD(D(p_{sr}(d), p_{ref}(d))) \tag{4}$$

with

$$D = \begin{bmatrix} p_{sr_1}^T \cdot p_{ref_1(1,3)}, 0, 0, 0, (-p_{sr_1}^T) \cdot p_{ref_1(1,1)} \\ 0, 0, 0, p_{sr_1}^T \cdot p_{ref_1(1,3)}, (-p_{sr_1}^T) \cdot p_{ref_1(1,2)} \\ \vdots \\ p_{sr_N}^T \cdot p_{ref_N(1,3)}, 0, 0, 0, (-p_{sr_N}^T) \cdot p_{ref_N(1,1)} \\ 0, 0, 0, p_{sr_N}^T \cdot p_{ref_N(1,3)}, (-p_{sr_N}^T) \cdot p_{ref_N(1,2)} \end{bmatrix} \quad (5)$$

and N being the number of feature points. The last column of matrix V represents the solution vector $h = V(:, 9) \in \mathbb{R}^{9x1}$, which gives the coefficients of the homography $H(d)$ in the form of

$$H(d) = [h_{11}h_{12}h_{13}h_{21}h_{22}h_{23}h_{31}h_{32}h_{33}]. \quad (6)$$

After the SVD a refinement step f_r given by

$$H_r(d) \quad = \quad f_r(H(d)), \quad (7)$$

to optimize the results and get the homography $H_r(d)$ is used.

After this step the homographs for the seven defined distances are available, but the distances in between are still missing. For this reason an interpolation step was done to determine a polynomial function of the degree 4 to approximate the homography of each position in the distance between 1m - 4m. All the homographs calculated in 7 for the distances $d \in M$, are used to calculate the coefficient vector $C \in \mathbb{R}^{5x1}$. The polynomial curve fitting function f_p is used to calculate the vector C for each element of the homography $H = (h_{i,j})_{i,j=1..3}$ with

$$C(h(i, j)) = f_p(H_r(d, i, j)) \ \forall \ d \in M. \quad (8)$$

Now, for a certain distance d_n all elements of the vector C are used to calculate with

$$H_n(d_n, i, j) = C(h(i, j)_1) \cdot d_n^4 + C(h(i, j)_2) \cdot d_n^3 + C(h(i, j)_3) \cdot d_n^2 + C(h(i, j)_4) \cdot d_n + C(h(i, j)_5) \quad (9)$$

the elements of a new homography H_n. The next section presents the test of the homography calculation process.

3.3.3 Testing Registration

For checking the accuracy of the homographs in the distances $d_n \in M_n := \{1.25, 1.75, 2.25, 2.75, 3.25, 3.75\}$ (in meters), the coefficient vectors C described in equation 9 are used. In table 2 the displacement of the calculated pixel positions in relation to the real measured pixel positions in x- and y-direction are shown. The average pixel error in x- and y-direction is less than 2 pixels, which is a promising result and the reason why using this approach for further tests is reasonable.

distance	avg pix err x	avg pix err y
1.25m	0.83	1.50
1.75m	0.67	0.67
2.25m	0.67	1.67
2.75m	0.17	0.67
3.25m	0.67	0.83
3.75m	1.67	1.33

Table 2. Accuracy and displacement of the calculated pixel positions in relation to the real measured pixel positions in x- and y-direction

4. Experimental Results

For the test of the novel evaluation method for silicon retina stereo sensors, three different test cases are used, to demonstrate and present the performance of the new method in comparison to the approach used till now. In all test cases the stereo sensors are static, without movement, and observe a dynamic scene.

In figure 4 all three test cases are shown. Figure 4(a) shows a laboratory scenario where a planar disc with a printed pattern is rotating at a fixed distance (1.5m) in front of the stereo sensors. In figure 4(b), the torso of a human body is shown, where a distance range of 1.1m from the hands till 1.7m of the head is covered, which represents a more closer real world scenario than the rotating disc. Figure 4(c) shows two persons walking around in 2m and 4m distance. This scenario was chosen to have a real world scenario with different objects at different distances, and where the object shape is not represented by a plane.

For all three test cases the average distance error e_{avg} in meter was calculated. Therefore, all depth values $p_{sr} = (x, y)^T$ of the depth map DM_{sr} generated by the silicon retina based stereo algorithm are processed with

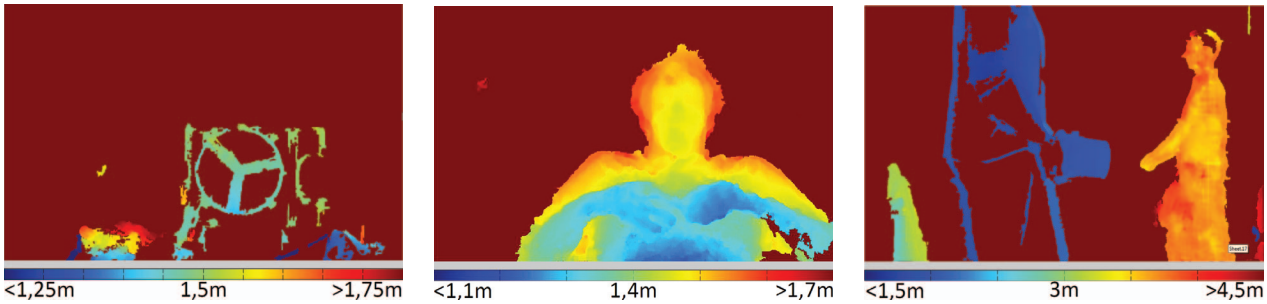
$$e_{avg} = \sum_{p_{sr} \in DM_{sr}} |DM_{sr}(p_{sr}) - DM_{ref}(H_n \cdot p_{sr})|, \quad (10)$$

whereby the depth values must fulfill the following constraint

$$\forall p_{sr} \in DM_{sr} | DM_{sr}(p_{sr}) \neq 0 \wedge DM_{ref}(H_n \cdot p_{sr}) \neq 0. \quad (11)$$

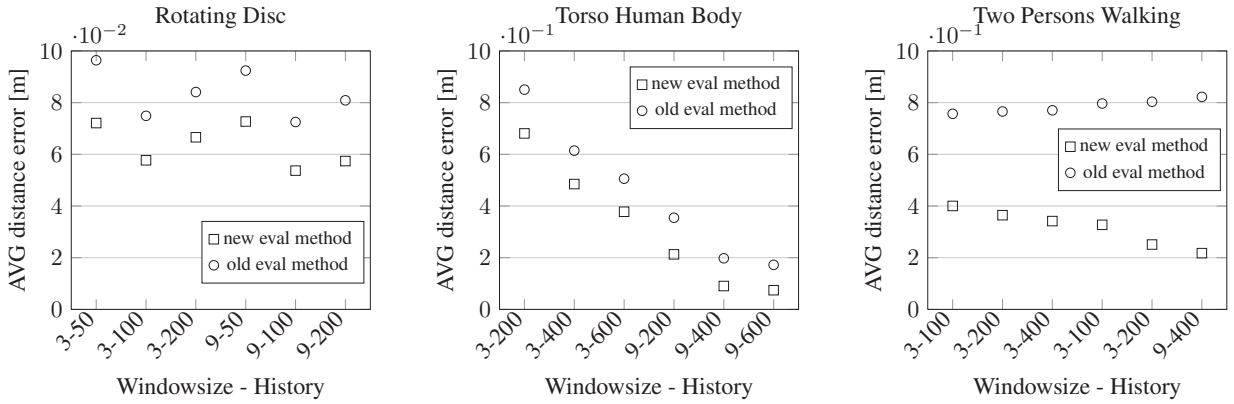
Here, DM_{ref} is the depth map used as ground truth data calculated by the reference stereo system.

The error was processed for different window sizes of the SAD stereo correlation and different history times (accumulation times) for grayscale image generation, of the silicon retina data stream. All results of the new approach were compared with the outcome of the old method, where all depth pixels were compared with a fixed distance, and are illustrated in figure 5.



(a) (b) (c)

Figure 4. Reference depth images from all three test cases; (a) planar rotating disc in 1.5m, (b) torso of a human body in a distance range of 1.1m-1.7m, (c) two independent persons (objects) in 2m and 4m.



(a) (b) (c)

Figure 5. Average distance error of all three test cases - (a) shows the results of the rotating disc in a fixed distance, (b) shows results of the human torso with different distances in the foreground, (c) shows the results of two persons walking around - Square marks in the plot show the results of the new approach in comparison to the circle marks which represent the outcome of the old evaluation approach.

Figure 5(a) shows the result of the rotating disc, where the average distance error between the old and new method was minimally increased by around 0.02m. In this case the improvement is not significant, because the rotating disc is represented by a single plane and also the old evaluation method is quite accurate. The usage of longer histories generate higher error rates because of the rotating disc, which moves on and more events are collected which do not represent the current position of the disc (motion blur).

The results of the second test case, the torso of a human body, are depicted in figure 5(b). In the case, where different distances are available, the improvement of the new method in contrast to the old one is much more significant. The old method evaluates all torso depth values at one pre-defined distance value, and therefore, the average error over all input datasets was increased by approximately 0.2m. Within this test case it is good to see that there is a continuous improvement of the data, if larger window sizes are used, and if the history of the event stream is increased to collect more events to match.

In the last test case, shown in figure 5(c), two persons in 2m and 4m distance were evaluated. Here, the old method performs worse because the evaluation of all depth values referred to one distance value. In this case we measured and compared all depth values of the scene with an artificial distance of 3m, which produces an average distance error of 0.8m with all different SAD results. In this real world scenario, the new approach achieves a much more accurate result. The measured average distance error was between 0.2m and 0.4m depending on the window size and event stream history. Furthermore, this test case shows also the outage of the old method on testing different algorithm settings. Different window sizes and histories have less or no influence on the output or show a behavior which is not related to the algorithm's performance. However, the new method facilitates a more exact evaluation of the stereo matching results, which visualizes the impact of different sets of parameters of a stereo algorithm applied on real world data. As a consequence, the new method is well suited for the development and evaluation of new stereo vi-

sion algorithms and their settings.

5. Conclusion

In this work we presented a novel approach to generate ground truth data for stereo vision systems based on silicon retina cameras. For conventional stereo systems, various ground truth datasets and methods to generate new datasets are available. Event-driven silicon retina based systems cannot be evaluated with these datasets and also specific datasets for retina sensors are not available. Up to now, only fixed objects at fixed distances were evaluated, but this does not represent real world scenarios were different objects at different distances along with various shapes and speeds are present. The introduced approach uses a video based stereo sensor with a sufficiently high accuracy and registers the stereo sensors' data onto each other with homographs. This method allows not only the evaluation of stereo matching results, but is also accurate enough to visualize the effect of different parameter sets. In all test cases the generated ground truth data provided accurate and comparable results. The results show that the new evaluation method with ground truth data enables the development of silicon retina based stereo matching algorithms and their comparison on real world datasets. In further research, new silicon retina based stereo matching algorithms will be developed and evaluated with the proposed approach. A different application of this method is also the fusion of 3D data, generated by different sensor technologies.

6. Acknowledgements

The project has been funded by the Austrian Security Research Program KIRAS - an initiative of the Austrian Federal Ministry for Transport, Innovation and Technology.

References

[1] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *IEEE Journal of Transactions on Pattern Analysis and Machine Intelligence*, 25:993–1008, 2003.

[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference CVPR*, Providence/USA, 2012.

[3] M. Humenberger, C. Zinner, M. Weber, W. Kubinger, and M. Vincze. A fast stereo matching algorithm suitable for embedded real-time systems. *Journal of Computer Vision and Image Understanding*, 114(11):1180 – 1202, 2010.

[4] J. Kogler, C. Sulzbachner, and M. Humenberger. Event-based stereo matching approaches for frameless address event stereo data. In *Proceedings of the 7th International Symposium on Visual Computing ISVC*, Las Vegas/USA, 2011.

[5] P. Lichtsteiner, J. Kramer, and T. Delbruck. Improved ON/OFF temporally differetiating address-event imager. In *Proceedings of the 11th IEEE International on Electronics, Circuits and Systems ICECS*, TelAviv/Israel, 2004.

[6] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128×128 120dB 30mW asynchronous vision sensor that responds to relative intensity change. In *Proceedings of the IEEE International Solid-State Circuits Conference ISSCC*, SanFrancisco/USA, 2006.

[7] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2), 2008.

[8] M. Mahowald. *VLSI analogs of neuronal visual processing: a synthesis of form and function*. Phd-thesis, California Institute of Technology, 1992.

[9] M. Mahowald and C. Mead. Silicon retina. *Jounal of Analog VLSI and Neural Systems*, pages 257–278, 1989.

[10] C. Mead and M. Mahowald. A silicon model of early visual processing. *Journal of Neural Networks*, 1(1):91–97, 1988.

[11] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Journal of Optical Engineering*, 51(02):021107, 2012.

[12] C. Posch, 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(1):259–275, 2011.

[13] P. Rogister, R. Benosman, S.-H. Ieng, P. Lichtsteiner, and T. Delbruck. Asynchronous event-based binocular stereo matching. *Journal of Neural Networks*, 23(2):347–353, 2012.

[14] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Journal of Computer Vision*, 47(1-3):7–42, 2002.

[15] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference CVPR*, pages 195–202, 2003.

[16] M. Sivilotti. *Wiring consideration in analog vlsi systems with application to field programmable networks*. Phd-thesis, California Institute of Technology, 1991.

[17] C. Sulzbachner, J. Kogler, and F. Eibensteiner. A novel verification approach for silicon retina stereo matching algorithms. In *Proceedings of the 52nd International Symposium Electronics in Marine ELMAR*, pages 467–470, Zadar/Croatia, 2010.

[18] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In *Proceedings of the 23rd International Conference Image and Vision Computing New Zealand IVCNZ*, pages 1–6, Christchurch/NewZealand, 2008.

[19] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the IEEE International Conference on Computer Vision ICCV*, pages 666–673, 1999.