Event-Driven Stereo Matching for Real-Time 3D Panoramic Vision

Stephan Schraml, Ahmed Nabil Belbachir
AIT Austrian Institute of Technology
Digital Safety & Security Department
New Sensor Technologies
Donau-City-Straße 1, 1220 Vienna, Austria
{stephan.schraml,nabil.belbachir}@ait.ac.at

Horst Bischof
Graz University of Technology
Institute for Computer Graphics and Vision
8010 Graz, Austria
bischof@icg.tugraz.at

Abstract

This paper presents a stereo matching approach for a novel multi-perspective panoramic stereo vision system, making use of asynchronous and non-simultaneous stereo imaging towards real-time 3D 360° vision. The method is designed for events representing the scenes visual contrast as a sparse visual code allowing the stereo reconstruction of high resolution panoramic views. We propose a novel cost measure for the stereo matching, which makes use of a similarity measure based on event distributions. Thus, the robustness to variations in event occurrences was increased. An evaluation of the proposed stereo method is presented using distance estimation of panoramic stereo views and ground truth data. Furthermore, our approach is compared to standard stereo methods applied on event-data. Results show that we obtain 3D reconstructions of 1024 × 3600 round views and outperform depth reconstruction accuracy of state-of-the-art methods on event data.

1. Introduction

The role of sparse information has been shown to be an effective technique for solving tasks or boosting the performance of many computer vision applications. Typical examples are highly condensed and possibly enriched descriptions of distinctive scene points used for e.g. solving visual correspondences, object matching and image analysis [1, 20, 32]. Stereo matching, which is recognized to be a computationally expensive task, can be efficiently computed using such sparse feature descriptors [28].

Another approach, providing a sparse representation of scene content, is given by the biologically inspired dynamic vision sensors (DVS) [16] that convey visual information by generating and passing events yielding a sparse visual code [3], which we call event-driven vision. These sensors allow high energy efficiency, high temporal resolution and compressed sensing [3] to efficiently solve computer vision tasks: e.g. real-time gesture control interface [14], event-driven formulation of epipolar constraint [4], stereo matching and 3D reconstruction [25, 6] and high-speed object classification [2].

By implementing the concept of sparse information at the pixel level, these sensors only detect scene content that has undergone a relative contrast change. Similar to biological neural systems, they feature massively parallel preprocessing of the visual information making use of an analog circuit at each pixel, combined with asynchronous event information encoding and communication. A scene content is thereby not presented as an image frame; instead the therein-detected relative intensity changes are coded and transferred as a variable stream of asynchronous events, efficiently capturing scene dynamics (frameless) at a high temporal resolution and high dynamic range. The advantages of event-driven vision include: Firstly, that static background containing redundant information does not lead to the generation of events, resulting in a massive reduction of data volume, which in turn helps to save processing time and computational resources. Secondly, relative contrast changes are mainly caused by moving objects, which eliminates the need for a time-consuming segmentation of moving (foreground) and static background objects as a preprocessing step, e.g. tracking of people [22].

Panoramic vision in 3D [13] has the advantage of providing a full 360° view; hence features and objects can be observed continuously supporting navigation [29] and localization tasks [11], e.g. a driverless google car. Laser-based systems [31] are capable of providing 360° panoramas in 3D. Although this technology guarantees high accuracy, it has a low vertical resolution at high cost (∼80k$). On the other hand, there have been a number of attempts to obtain panoramic views based on stereo vision in literature [10, 21, 30]. However, achieving the performance required for applications targeted on embedded systems in terms of real-time capability with limited computing re-

sources is still a challenge. Based on event-driven vision, a multi-perspective (rotating) stereo system consisting of a pair of line DVS, each having a resolution of 1024×1 pixels, was recently proposed [3]. This system is capable of providing high resolution stereo panoramic views of up to 10 pan/s. Although standard stereo methods can be applied to event views, they do not cope well with event-driven vision, since these methods were conceived for conventional images relying on prerequisites (e.g. dense images) that do not hold for event-driven vision. We therefore believe that tailored event-driven methods have to be developed in order to efficiently exploit the advantages of event-driven vision.

In this paper, we present an efficient stereo matching method for event-driven vision, introducing a novel cost measure for asynchronous events generated by a multi-perspective DVS system (Section 3). Our algorithm is able to exploit the temporal information of events generated from scene contrast as a matching criteria having the advantage that no direct images are required; Section 4. In order to evaluate our method, we present an experimental comparison with conventional state-of-the-art stereo methods on events and ground truth data for several scenes, demonstrating the advantages of event-driven vision in Section 5. Section 6 concludes the paper and highlights the main findings.

2. Related work on Event-Driven Stereo Vision

Since the early work on event-driven vision in [17], several prototypes of dynamic vision sensors [16, 5, 3] were developed in the last decade and numerous stereo matching approaches based on events were published.

Mahowald et al. [18] presented in 1989 a stereo matching method using static and dynamic image features. An area-based approach using an adapted cost measure for event data was presented by Schraml et al. [27] and demonstrated in a real-time tracking application. Later Kogler et al. [12] provided a comparison between area-based and feature based methods and Eibensteiner et al. [6] implemented in hardware a high performance event-driven matching approach, which is based on time-correlation. In the work from Rogister et al. [25] (two static DVS, simultaneous stereo vision) stereo matching is based on the idea that correlated events are likely to appear within a small time interval and on the same epipolar line. Small variations in timing are tolerated. The work of Piatkowska et al. [23] combines this approach with a dynamic network such that the history of events contributes to the stereo matching. Aside from this, in Lee et al. [14] a stereo system was used for gesture control. Although disparity was not explicitly calculated, the left and right data stream was combined so that fore- and background motion could be distinguished. These methods use data from two static DVS based on simultaneous stereo vision.

In order to get a better understanding of event-driven stereo matching we propose a subdivision of stereo methods into three groups based on how the correspondence is calculated: a) Classical stereo methods: Stereo matching is performed using a conventional stereo method based on area sensors. The events are transformed into an image-like representation, for example by integration of events. b) Event-driven methods from simultaneous DVS. Using the sensors' high temporal resolution, temporal coherence is used to find matching events between left and right sensor data. c) Event-driven matching from non-simultaneous DVS (our method). The disparity is given by the time difference between corresponding event pairs.

The majority of these methods is associated with groups a) or b), both using area DVS that simultaneously records the scene. These methods rely on event occurrence at a close timely distance between left and right sensor. However, data from group c) cannot be handled by these methods since left and right views are non-simultaneously recorded.

In more detail: In an event-driven vision system multiple events generated from one pixel extend in time. For the simultaneous recording of the left and right views, the data collection of an event-driven system based on area sensors, therefore always spans a 3D space (x-y-t). The pixel activations in the left and right view deviate in the spatial domain and are inferior in time domain. Either a spatial activation pattern (in x-y space with flattened time domain) or the temporal coherence based on statistics may be used for defining a correlation metric. However, in non-simultaneous stereo vision like in the case of concentric panoramas [15] data collection spans a two dimensional (x-t) space, where 2D spatial information is not explicitly available. Here, pixel activations in the left and right view always deviate in the time domain, i.e. corresponding events appear at a different time. The realization of an event-matching algorithm for handling non-simultaneous vision is an open issue.

We solve the problem of finding corresponding events in the time domain by defining a novel cost measure that is based on event distributions using inter-event distances, which is tolerant to variations in time as well as in the number of total events.

3. Panoramic Stereo Vision Based on Biologically Inspired Dynamic Vision Sensors

This section briefly describes the event representation and the key characteristics of the multi-perspective stereo system using DVS. The system consists of a rotating pair of dynamic vision line sensors V_1 and V_2 arranged symmetrically at an equal distance R to the rotational center C_0, generating symmetric pairs of concentric panoramas (Figure 1). A multi-perspective panoramic view is acquired by collecting the stream of asynchronous events during the system revolution, yielding a sparse visual code. Due to the

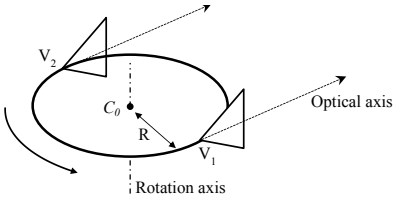

Figure 1. 360° Panoramic stereo vision setup.

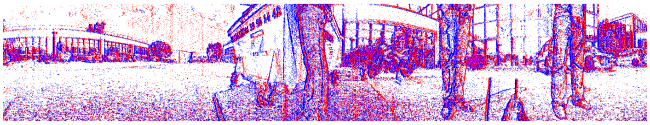

Figure 2. Events from a panoramic view showing the events polarity; ON-events (red) and OFF-events (blue).

symmetric setup, the ratio of the vertical scaling factor between the left and right view $s_{2:1} = 1$ [15]. Unlike stereo matching on standard images, where the disparity quantifies the parallax effect of a stereoscopic view by means of Euclidean distance, this system measures disparity as a time difference dt between observation of a scene point by the first and second sensor. The depth Z can be formulated for sensors that have their optical axis perpendicular to the swing line C_0V_{12} as:

$$Z = \frac{R}{\sin(\frac{\delta}{2})} \tag{1}$$

where R is the radius. By using the system's revolution rps (assumed constant), the measured disparity given as an angle is easily calculated as $\delta = 360° * rps * dt$. The asynchronous events contain three elements: (1) the address (position) y of the triggered pixel, (2) the time of occurrence t and (3) the polarity p of the relative contrast change, coded as (-1, +1) reflecting a negative (OFF-event) or positive (ON-event) relative contrast change respectively. We use the notation $e^j(y, p, t)$ for a single event, where the superscript j indexes the sensor that generated this event. The example in Figure 2 shows the event polarity information.

4. Stereo Matching

In this section we describe the individual processing steps for event-driven stereo matching for asynchronous vision. Stereo matching refers generally to the search for corresponding primitive descriptions in views from different perspectives [19] in order to reconstruct 3D information by triangulation. In event-driven vision, it has to be considered that single events do not carry enough information to be matched thoroughly. Although events can be transformed into a map representation, the common solution of reducing matching ambiguity by defining areas as matching primitives does not give satisfactory results for two reasons: i)

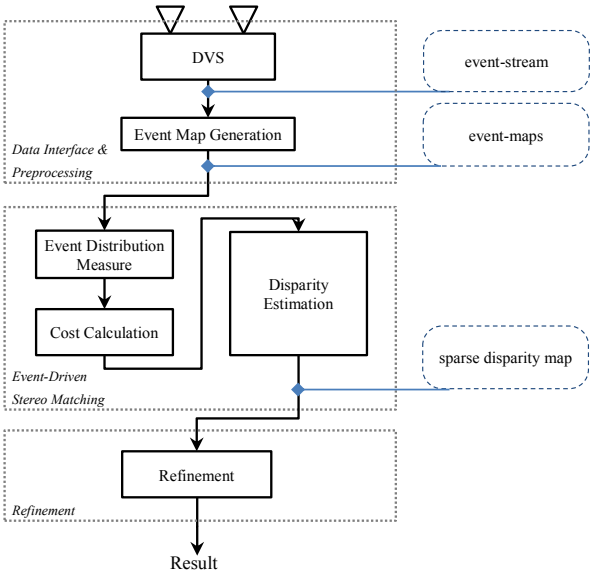

Figure 3. Event-driven stereo algorithm workflow.

Event data is sparsely distributed, so that there exist positions in the transformed event map where no data is available and hence cannot be matched. ii) Events do not appear at the same time, meaning that sequences of events from the left and right view are not identical. The direct matching of events would therefore be inappropriate.

The stereo matching in our approach (two line DVS, non-simultaneous stereo vision) is based on the idea that correlated sequences of events are likely to have a similar event-count and inter-event timing. Variations in the timing and the number of generated events can be quantified by the proposed cost measure. For the purpose of finding corresponding event positions, a novel event-driven stereo matching algorithm was developed. This algorithm can handle the absence of information and cope with inter-event sequence variations by defining a matching method based on the distribution of events rather than on pixel wise correspondence. The proposed stereo algorithm is outlined in Figure 3 and consists of three steps.

4.1. Data Interface & Pre-processing

The purpose of this module is to acquire event data from both sensors and to prepare data for the event-driven stereo matching step. Event-data is thereby transformed so that spatial context is restored.

4.1.1 Event Map Generation

Stereo reconstruction requires multiple views of the same scene to build a depth map. Typically, spatial context is used in stereo matching. As spatial context is not evident in event streams, it has to be restored. The first step is to transform

Figure 4. Event map from outdoor scene (without coding the polarity).

the timestamps t of events into image coordinates x using:

$$x(t) = c_x * ((t * rps) \, mod \, 1) \tag{2}$$

Variable c_x determines the image horizontal resolution, which can be interpreted as a horizontal scaling factor. If resolution c_x is set to the reciprocal of sensors's timestamp resolution then the event map is built without time quantization and thus without loss in horizontal resolution. In our example the sensor system timestamp period was set to 40 μs, which is equivalent to 25000 columns. The term $((t * rps) \, mod \, 1)$ corresponds to the angle, represented as a fraction of one revolution. The second step is to render the panoramic event map $E(x, y)$ by accumulating the event polarities:

$$E^j(x, y) = \sum_{e_i^j \in M(t_1, t_2), y_i = y} (p_i \,|\, x_i(t_i) = x) \tag{3}$$

where y_i and p_i are the spatial location and polarity of the event e_i and (x, y) are the image coordinates. Such a reconstructed event map can be viewed as a panoramic image, showing contrast change at high-resolution, (Figure 4).

4.2. Event-Driven Stereo Matching

4.2.1 Event Distribution Measure

An ordered list of events generated from a sensor's pixel form an event sequence. Two correlated sequences of events taken from the same scene at two different times or recorded from two different sensors show similar distribution, however they do not exactly match for several reasons. These are caused by internal factors e.g. timing jitter resulting from fabrication tolerances of sensor circuits and external factors, like reflectance variances or changes of perspective. Considering this, a novel cost measure is developed that is based on measuring the relative timing differences of two event sequences (one-dimensional patches) from the left and right sensor lines: the positions of events (time of occurrence) in the first sequence are compared with those of the second by means of minimal distances between event positions. The cost is equivalent to the sum of all these distances. The principle and examples for a low cost and a high cost measure are illustrated in Figure 5.

In a first step, we define a measure that efficiently describes the local distribution of event positions within a single event-sequence, the *Non-Zero-Distance* (NZD) function. Such event-sequences are found in the transformed

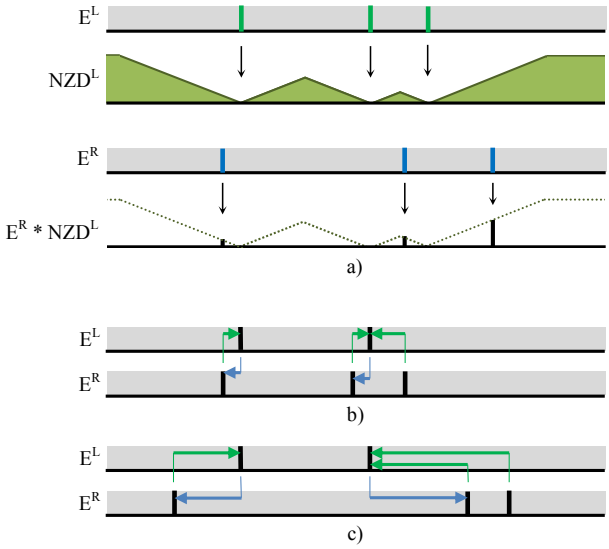

Figure 5. Similarity measure. Partial cost calculation based on minimal distances between event positions from the left to right (blue) and the right to left (green) event maps. Event positions are indicated by vertical bars. Illustration of cost calculation (a). Example from a low cost case (b) and a high cost case from different event-sequences.

event-map as segments of horizontal lines. The NZD assigns to each point in the event map $E(x, y)$ the minimum distance to any other event of the same image row (i.e. generated by the same sensor pixel), which can be formulated by:

$$NZD^j(x, y) = \min_i (abs(x - i) \,|\, E^j(i, y) \neq 0) \tag{4}$$

where j indices the sensor.

4.2.2 Cost Calculation

The cost measure is calculated as follows: The partial cost using segments of E^L and NZD^R, where L, R denote the left and right sensor, calculates to:

$$C^{L,R}(x, y, d) = \sum_{w \in W} E^L(x+w, y) * NZD^R(x+w+d, y) \tag{5}$$

where W is the patch window size. In order to maintain symmetry, partial cost is also similarly calculated with exchanged E and NZD:

$$\hat{C}^{L,R}(x, y, d) = \sum_{w \in W} NZD^L(x+w, y) * E^R(x+w+d, y) \tag{6}$$

It should be mentioned that the two partial costs have the same sensor, indexed by L, as reference and are thus different to cross-check, which is characterized by exchanging

the role of the reference and the dependent sensor. This technique is applied during disparity estimation. Finally, the total cost of a match is modeled as the sum of partial costs:

$$C = \begin{cases} C^{L,R} + \hat{C}^{L,R}, & \text{if } (n^L \geq \tau) \wedge (n^R \geq \tau) \\ C_{max}, & \text{else} \end{cases} \quad (7)$$

with $n = \sum_{w \in W} E(x, y)$ the number of events in a segment and τ the minimum event count. If a segment do not contain enough events, partial cost calculation is omitted and set to a maximum value. For evaluation, the following settings were used throughout all experiments in this paper: minimum number of $\tau = 3$ events and maximum cost $C_{max} = W$, which empirically performed well.

4.2.3 Disparity estimation

In the previous step a cost matrix, the disparity space image (DSI) which is sparsely filled, was built. In order to find the optimal method several disparity computation strategies were evaluated: a winner-takes-all (WTA) strategy, the semi-global matching method [8] and a dynamic-programming (DP) approach. This DP approach delivered experimentally the best results, and was therefore used in the algorithm. Disparity thereby results from computing the minimal cost path through the cost map of all pairwise costs between two corresponding event map lines [26].

4.3. Refinement

The *refinement* step is used to filter small outliers and streak effects, which may remain from a 1D search in the disparity map. First, the disparity map of the right event map is obtained and a cross-check is applied to detect unreliable matches. A match is thereby accepted if both disparities coincide; otherwise, it is discarded. Consecutively, the disparity map is refined by replacing disparities that are very different to its surroundings with a locally mean value, following the approach from [8]. This method allows most of the outliers to be eliminated. Finally, the disparity of event positions that could not be matched, is estimated as a locally smooth interpolation to valid neighbors.

4.4. Standard Stereo Matching

We compare our method in Subsection 5.2 with state-of-the-art stereo methods. The stereo matching of event-data is performed on transformed event-maps. As these methods provide dense disparity maps, we mask the disparity map with the event map to get a sparse map similar to the event-driven stereo results.

5. Evaluation

This section provides an evaluation of the proposed event-driven stereo matching using distance measures from panoramic views and the results from state-of-the-art stereo methods applied to transformed event-data.

All event-data were recorded by the multi-perspective DVS system with identical settings. The optical focal length used was $f = 4.5$ mm, resulting in a vertical field of view of 45°. In this setup the sensor resolution is 1024×1 pixels, while the reconstructed event-map resolution is of variable size $1024 \times c_x$. The photosensitive size is 12 μm in a horizontal direction, which is projected through optics to approximately 0.152°. A panorama may therefore be represented by $\frac{360}{0.152} = 2368$ pixels so that projected field of view of the pixels does not overlap but covers the entire panorama. This can be seen as the native image resolution. However, since the sensor's timing resolution is much higher and multiple events may occur from the detection of strong contrast changes, we also investigated the matching performance of higher resolution depth maps up to 1024×3600 pixels. Since the image resolution of transformed event maps also affects the sparseness of data represented therein, we used smaller image resolutions as well in order to evaluate the performance of the stereo methods with regard to data sparseness. If an event stream is mapped to a higher image resolution, clearly the events become less cluttered, i.e. the events are more distributed and a larger quantity of image area remains empty. For our evaluation we therefore use image resolutions of 256×700, 256×1400, 512×1400, 1024×2800 and 1024×3600 pixels. The event maps are built according to Equation 3.

5.1. Panoramic Depth Estimation

In our first evaluation we compared panoramic depth map reconstruction converted to a Cartesian coordinate system with ground truth data. Several recordings were performed with objects of various shapes and sizes placed at different distances. A camera image of one test set is shown in Figure 6. This image is shown for illustration purposes only and therefore does not claim an exact alignment to the recorded event data. The true distances to each object were measured manually by a laser distance meter. We performed stereo matching using the proposed method of these test sets followed by a transformation of resulting event disparities in metric distances (Equation 1). For a good alignment of distance measures we shifted disparities by -0.65 degrees, which allows an inclination of the two sensors to be corrected. Results are presented in Figure 7 with color coded distances. Objects can be recognized by their shape and the unique distance coding of their boundaries. The books in the middle image (B) also show well-matched inter-object structures.

In Figure 8 the same data is represented as a 2D map

Figure 6. Camera image of test recording (Figure 7)(B) taken with a mobile device.

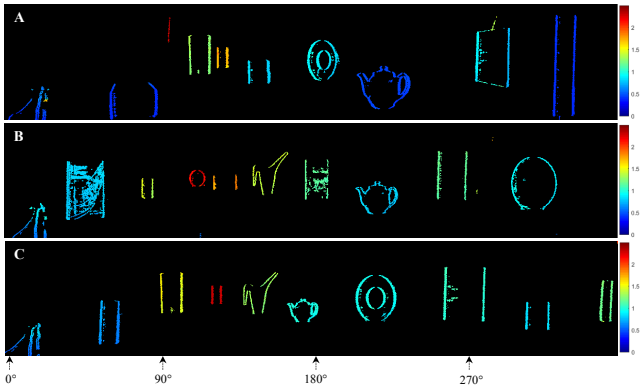

Figure 7. Experimental results of test recordings showing the distance color coded.

by registering the events using their horizontal position and distance estimation from a top down view in a Cartesian coordinate system. Note that data from all sensor pixels are used. To eliminate the stairs effect from discrete disparity values, the calculated distances of each event are scattered within one pixel disparity tolerance. The concentric circles indicate the depth resolution of single disparity steps. Ground truth measurements are plotted as red crosses. We can see that most of the edges are within a range of one or two disparities, indicating the total variation of measured event distances of these objects edges. It can be noticed that the distance estimation thereby matches the ground truth. Thanks to the high vertical and dense resolution, an object is detected by a significant number (in our examples depending on size and distance 200-500) of pixels.

5.2. Standard Stereo Matching using Transformed Event-Data

We compared our event-driven stereo matching (EDS) algorithm with two state-of-the-art local stereo methods based on transformed event-data: The "Fast Cost-Volume Filtering" (FCVF) [9] method makes use of a filtering technique based on an edge preserving filter, which seems adequate for handling event data. This method is a top performer in the Middlebury database [26]. The "Libelas" method [7] estimates disparity by forming triangulation on sparsely supported points and uses this to efficiently exploit the disparity search space. An implementation for both is publicly shared.

For this test we used recordings from a structured light

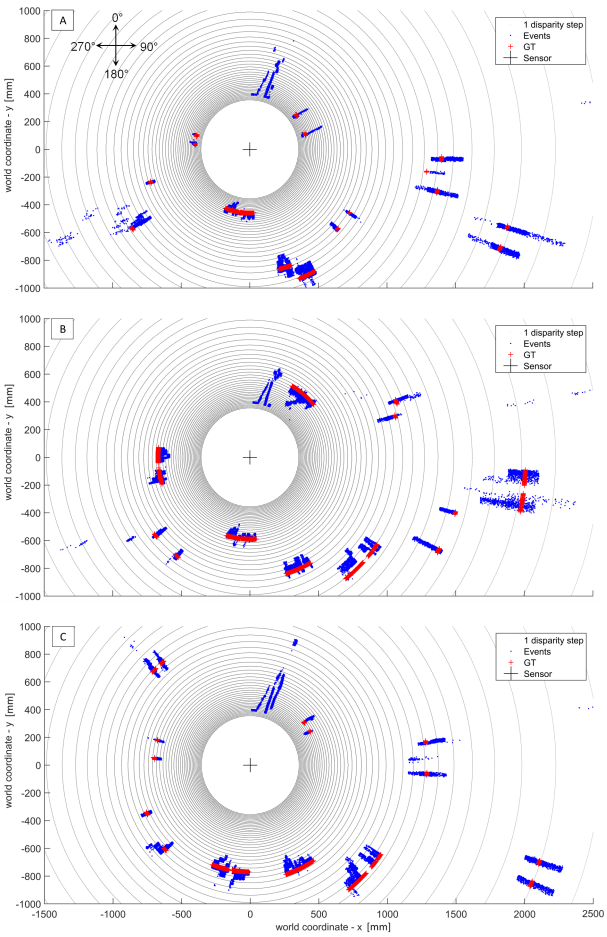

Figure 8. Experimental results of event-driven stereo transformed to Cartesian map compared with ground truth.

system as ground truth data, which is masked with the reference (=left) event map. As an error measuring metric we use the error rate, defined as the fraction of incorrect disparities according to a threshold, which were set to $t = 1.0$ and $t = 2.0$. Sparse event information means that a typically large fraction of the disparity map remains empty, and as such should be excluded from evaluation. Therefore, pixels containing no event information are not considered when calculating the error rate. The comparison for two examples is presented in Figure 9.

Note that those large unstructured areas in the background, which are for image-based methods typically difficult to match and therefore more likely to contain mismatches, are not considered in event-driven stereo since these areas do not generate events. Results are listed in Table 1. The *pixel* column indicates the sparseness as the fraction of pixels in the event map containing an event. While FCVF and Libelas show a tendency toward higher error rates with increasing image resolution, i.e. sparseness, our

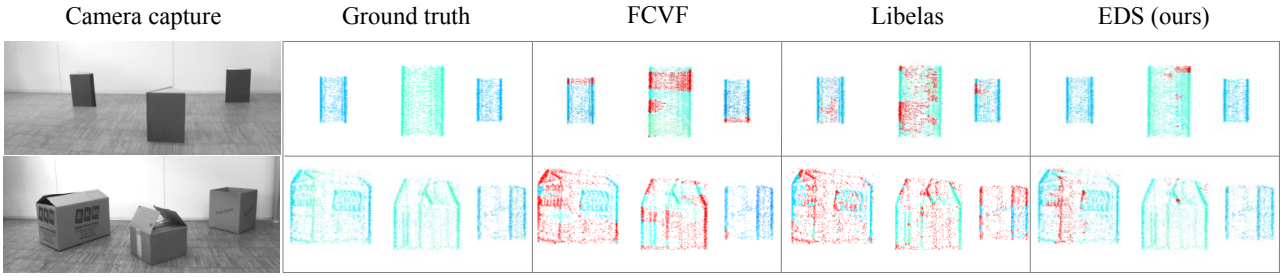

Figure 9. Results on a subset of the test recordings with ground truth. Note that the camera capture was taken from a different (elevated) position. Ground truth comprises only events from foreground objects. The background as well as the floor is not considered in evaluation. Errors ($t = 2.0$) are plotted in red.

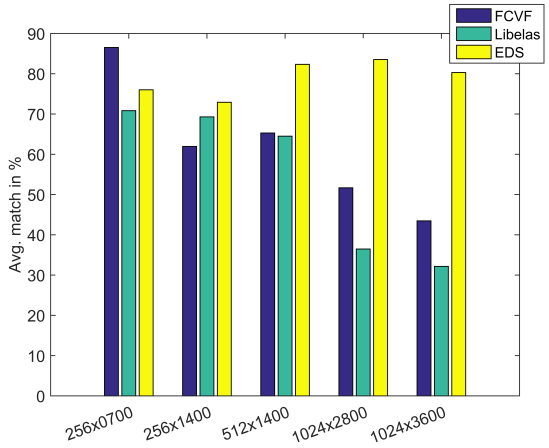

Figure 10. Average match performance vs. image resolution. With higher image resolution data becomes sparser. Our EDS method achieves a high matching performance for all image resolutions and outperforms standard stereo methods, particularly for high image resolutions.

method delivers good results among all image resolutions and the highest accuracy in all test cases for (native) image resolution of 1024×2800 and above. The average matching performance (error<1.0) stresses this context, as in Figure 10 presented. It is interesting that although our method was designed for sparse data, it still performs competitively and in three of six cases performs best for the lowest evaluated image resolution, where pixel density is above 50%.

5.3. Real-World Scenario

This section demonstrates the success of the proposed method on a challenging real-world scenario. Using the same system, a scene was recorded outdoor at 10 rps under bright sunlight conditions including highly textured areas like gravel path, grass, bushes and trees. The corresponding left event-map is shown in Figure 4. The result is presented in Figure 11 with color-coded distances, showing that the

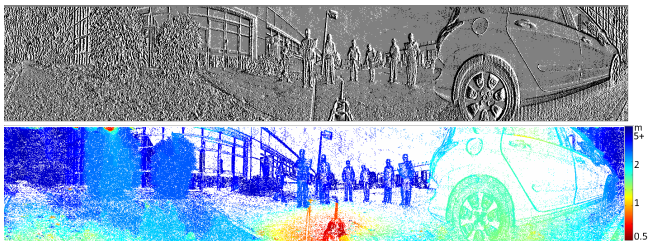

Figure 11. Experimental result of a real-world scenario. Transformed event-map (top) and stereo result (bottom).

people and the building in the rear as well as the car and the highly textured bushes are matched well, while the very slanted gravel path is problematic.

5.4. Discussion

The performance of proposed event-driven stereo matching has been demonstrated on three examples including accuracy measurement, comparison to standard stereo methods and real-world results. Our approach aims at real-time 3D reconstruction, which is therefore based on local operations without the need for global optimization; alhough global optimization methods - like [24] - can also generate decent results. However, global methods are contradictory to the concept of sparse and non-simultaneous event-driven vision. We therefore found it misleading to have a mix in the concept of how stereo matching is performed and restricted the evaluation to FCVF [9] and Libelas [7], which are both fast local methods. In our approach, the event-driven stereo matching may be processed for each line in the disparity map in parallel.

We have implemented the algorithm in C# for real-time performance, which is achieved for half native resolution, i.e. 1024×1400, by now using an un-optimized code. The cost calculation requiring ∼2/3 of total computation time is well-suited for optimization. Moreover, the computation cost is linear in scale with image resolution.

The robustness and limitations of the proposed method

Image	size	pixel	FCVF t=1.0	Libelas t=1.0	EDS t=1.0	FCVF t=2.0	Libelas t=2.0	EDS t=2.0
im1	256*700	0.57	**1.5**	45.4	31.0	**0.7**	43.3	10.6
im1	256*1400	0.42	13.3	30.6	**11.6**	9.9	28.0	**2.9**
im1	512*1400	0.31	8.1	44.5	**6.6**	4.9	42.7	**0.8**
im1	*1024*2800*	0.12	34.6	84.6	**10.6**	27.8	78.6	**3.0**
im1	1024*3600	0.09	49.6	72.4	**22.0**	34.9	63.3	**3.4**
im2	256*700	0.58	**8.5**	21.4	24.8	3.5	18.8	**0.3**
im2	256*1400	0.44	58.2	**14.2**	48.8	48.2	5.8	**5.7**
im2	512*1400	0.31	52.3	**22.5**	28.6	44.4	19.3	**2.9**
im2	*1024*2800*	0.12	35.8	41.2	**7.0**	35.8	24.1	**3.5**
im2	1024*3600	0.10	50.4	44.7	**3.5**	41.1	28.8	**1.0**
im3	256*700	0.56	30.3	20.8	**16.2**	10.4	16.4	**1.7**
im3	256*1400	0.41	42.6	47.3	**20.8**	31.1	40.1	**5.0**
im3	512*1400	0.30	43.7	39.5	**17.7**	25.0	34.3	**2.8**
im3	*1024*2800*	0.11	74.6	64.7	**31.8**	57.5	51.4	**10.2**
im3	1024*3600	0.09	69.6	86.4	**33.7**	45.1	78.9	**13.2**

Table 1. Errors for each of the evaluated methods. Our EDS method yields the highest accuracy for most of the test cases. In particular, for stereo reconstructions of higher image resolutions - more sparseness - and reconstructions of native image resolution, our method performs best for both error thresholds.

rely on the capability and sensitivity of both sensor lines to capture edges in natural scenes. This is compromised by heavy noise or low event generation due to weak contrast changes. While noise can stem from increased sensor sensitivity, both noise and low event generation can occur in low lighting situations, when the limits of the high dynamic range of the sensor are reached. Since an edge is typically represented by several events, a few noisy events are tolerated very well. However, the NZD function gets flawed with an increasing number of noise events. The challenge of the method is to adapt the choice of the sequence size to be large enough to cover all events from an edge but as small as is possible to reduce the number of (random) noise events within a sequence. A weak contrast results in large timing variations, i.e. generates widely scattered events that cannot be matched. The disparity of these events are estimated in the refinement step.

6. Conclusion

We have presented a novel stereo matching method for non-simultaneous event-driven vision. It exploits the sparse event information as a result of scene contrast as a matching criteria for efficient 3D reconstruction in real-time out of 360° panoramic views. We have compared our method with ground truth and with two state-of-the-art stereo matching methods, which were designed for standard camera images. Results showed that the stereo reconstruction of scene contrasts detected at various distances agree with ground truth data in Cartesian map representation. The experiments revealed that thanks to the novel cost measure our tailored event-driven stereo method accurately reconstructs 3D information of event-data over a wide range of sparseness. It outperforms standard state-of-the-art stereo methods on sparse event-data, particularly for high resolution panoramic images. Results on the natural scene show the usability of the method and the capability of the method for application in a natural environment. In future we plan to investigate the properties of this matching method on more recordings of natural environments with varying environmental conditions.

Acknowledgement

This work is supported by the project BiCa360° (grant number 835925) from the Austrian Research Promotion Agency.

References

[1] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2012.

[2] A. Belbachir, M. Hofstätter, M. Litzenberger, and P. Schön. High-speed embedded-object analysis using a dual-line timed-address-event temporal-contrast vision sensor. *IEEE Trans. on Industrial Electronics*, 58(3):770–783, Mar. 2011.

[3] A. Belbachir, S. Schraml, M. Mayerhofer, and M. Hofstätter. A novel hdr depth camera for real-time 3d 360-degree panoramic vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 419–426, June 2014.

[4] R. Benosman, I. Sio-Hoi, P. Rogister, and C. Posch. Asynchronous event-based hebbian epipolar geometry.

IEEE Transactions on Networks and Learning Systems, 22(11):1723–1734, 2011.

[5] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch. Activity-driven, event-based vision sensors. *IEEE International Symposium on Circuits and Systems ISCAS*, pages 2426–2429, May 2010.

[6] F. Eibensteiner, J. Kogler, and J. Scharinger. A high-performance hardware architecture for a frameless stereo vision algorithm implemented on a fpga platform. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 637–644, 2014.

[7] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[8] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328341, 2008.

[9] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(2):504 – 511, 2013.

[10] W. Jiang, M. Okutomi, and S. Sugimoto. Panoramic 3d reconstruction using rotational stereo camera with simple epipolar constraints. *Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 2006.

[11] J. Kim, K.-J. Yoon, J.-S. Kim, and I. Kweon. Visual slam by single-camera catadioptric stereo. *SICE-ICASE, International Joint Conference*, 2006.

[12] J. Kogler, C. Sulzbachner, and W. Kubinger. Bio-inspired stereo vision system with silicon retina imagers. *Lecture Notes in Computer Science Volume 5815, 2009, pp 174-183*, 5815:174–183, 2009.

[13] H. Koyasu, J. Miura, and Y. Shirai. Realtime omnidirectional stereo for obstacle detection and tracking in dynamic environments. *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 31–36, 2001.

[14] J. H. Lee, T. Delbrück, M. Pfeiffer, P. K. Park, C. W. Shin, H. Ryu, and B. C. Kang. Real-time gesture interface based on event-driven processing from stereo silicon retinas. *IEEE Trans. on Neural Networks and Learning Systems*, 2014.

[15] Y. Li, H. Shum, C. Tang, and R. Szeliski. Stereo reconstruction from multiperspective panoramas. *IEEE TPAMI*, 26(1):45–62, 2004.

[16] P. Lichtsteiner, J. Kramer, and T. Delbrück. Improved on/off temporally differentiating address-event imager. *11th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2004)*, pages 211–214, May 2004.

[17] M. Mahowald. Vlsi analogs of neuronal visual processing: a synthesis of form and function, ph.d. dissertation. *California Institute of Technology*, 1992.

[18] M. Mahowald and T. Delbrück. Cooperative stereo matching using static and dynamic image features. *Analog VLSI Implementation of Neural Systems*, 80:213–238, 1989.

[19] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283287, 1976.

[20] B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

[21] S. Peleg, Y. Pritch, and M. Ben-Ezra. Cameras for stereo panoramic imaging. *IEEE Conference on Computer Vision and Pattern Recognition*, 1:208214, 2000.

[22] E. Piatkowska, A. Belbachir, S. Schraml, and M. Gelautz. Spatiotemporal multiple persons tracking using dynamic vision sensor. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 35–40, June 2012.

[23] E. Piatkowska, A. N. Belbachir, and M. Gelautz. Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013.

[24] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM Journal on Imaging Sciences*, 3(4):1122 – 1145, 2010.

[25] P. Rogister, R. Benosman, I. Sio-Hoi, and P. Lichtsteiner. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Networks and Learning Systems*, 23(2), 2012.

[26] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Intl J. Comput. Vision*, 47(1):742, 2002.

[27] S. Schraml, A. Belbachir, N. Milosevic, and P. Schön. Dynamic stereo vision system for real-time tracking. *IEEE International Symposium on Circuits and Systems*, pages 1409–1412, 2010.

[28] S. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1589, 2014.

[29] W. Stürzl and H. A. Mallot. Vision-based homing with a panoramic stereo sensor. *Biologically Motivated Computer Vision Lecture Notes in Computer Science*, 2525(1):620–628, 2002.

[30] T. Svoboda and T. Pajdla. Panoramic cameras for 3d computation. *Czech Pattern Recognition Workshop*, February 24(1):63–70, 2000.

[31] Velodyne Lidar Inc. HDL High Definition Lidar, 2013.

[32] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and Y. Shuicheng. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.