Asynchronous Event-Based 3D Reconstruction From Neuromorphic Retinas

João Carneiro[a], Sio-Hoi Ieng[b], Christoph Posch[c], Ryad Benosman[d]

Université de Pierre et Marie Curie - Institut de la Vision, 17 rue Moreau, 75012 Paris, France

[a]*joao.carneiro@inserm.fr, Phone:+33 1 53 46 22 66*
[b]*sio-hoi.ieng@upmc.fr, Phone:+33 1 53 46 26 79*
[c]*christoph.posch@inserm.fr, Phone:+33 1 53 46 26 79*
[d]*ryad.benosman@upmc.fr, Phone:+33 1 53 46 26 77*

Abstract

This paper presents a novel N-ocular 3D reconstruction algorithm for event-based vision data from bio-inspired artificial retina sensors. Artificial restinas capture visual information asynchronously and encode it into streams of asynchronous spike-like pulse signals carrying information on e.g. temporal contrast events in the scene. The precise time of the occurence of these visual features are implicitly encoded in the spike timings. Due to the high temporal resolution of the asynchronous visual information acquisition, the output of these sensors is ideally suited for dynamic 3D reconstruction. The presented technique takes full benefit of the event-driven operation, i.e. events are processed individually at the moment they arrive. This strategy allows to preserve the original dynamics of the scene, hence allowing for more robust 3D reconstructions. As opposed to existing techniques, this algorithm is based on geometric and time constraints alone, making it particularly simple to implement and largely linear.

Keywords: Stereo vision, neuromorphic vision, asynchronous event-based vision, 3D reconstruction

1. Introduction

Stereovision is the ability to extract depth from disparities of overlapping views. The fact that various animals rely on stereopsis as one of the mechanisms to perceive depth has been acknowledged for centuries now. Nevertheless depth perception from multiple views has always been an open research topic for both biology and machine vision. In computer vision, potential applications ranging from tele-immersion to robot navigation require users to be able to continuously produce 3D models of their surrounding environment in real-time.

State-of-the-art artificial vision systems rely on frame-based acquisition of the visual information. This acquisition strategy is not able to convey the temporal dynamics of most natural scenes and, additionally, produces large amounts of redundant data. Due to these fundamental weaknesses of current visual data acquisition, even the latest developments in stereo computer vision are still far from reaching the performance of comparatively "simple" and small biological vision systems.

Neuromorphic silicon retinas are solid-state vision sensors that mimic the behaviour of biological retinas, asynchronously encoding visual signals pixel-individually and usually at high temporal resolution. The usage of these recently developed devices in stereovision systems enables us to rethink the current approaches to the correspondence problem, supporting the development of spike-based, bio-inspired vision algorithms closely related to neurophysiological models.

In this paper we present an event-based trinocular stereo matching and reconstruction algorithm for event-based vision data. We use the properties of silicon retina vision sensors, such as high temporal resolution and response to relative light intensity changes, to address the stereo matching problem and produce high performance 3D reconstructions of visual scenes by applying well-known epipolar geometry in an event-based fashion. Furthermore we show that the trinocular and temporal constraints alone are insufficient to ensure a unique solution to the stereo correspondence problem and provide a bayesian inference method for discarding incorrect matches.

1.1. Stereo vision and depth perception

The stereovision mechanism has been acknowledged as a method for depth perception for centuries with first theories being proposed by Descartes and Newton [1]. The activity of cortical neurons responding to depth was however recorded for the first time in 1967 [2]. Since then, several studies aiming at understanding the neurophysiological basis of stereopsis have been developed, successfully showing the cortical response to binocular disparities [1]. To extract depth from stereovision, our brain examines disparities on the two retinal images of an object. To achieve this, it must first be able to determine correspondences between points in both views. This is known as the stereo correspondence problem. Several models to explain how the brain solves the stereo correspondence problem have been proposed with Parvati [3] and Jeremiah [4] in 1975 amongst the first.

Although humans seem to perform stereo correspondence effortlessly, the problem is still ill-posed since scientists struggle to reproduce a convincing model on a machine. Our brain uses complex cues from the outside world and from knowledge gained through experience to impose additional constraints (e.g. color, opacity, spatial and temporal coherence) in order to solve the stereo matching problem [5].

1.2. Stereo vision in computer vision

The depth estimation problem from multiple image views is one of computer vision's Holy Grail. The field of machine vision knows a wide variety of of stereovision algorithms which can be coarsly classified into two categories:

- feature-based techniques which consist in matching feature points across the images,
- area-based techniques which use image domain similarity metrices for the matching operation.

In [6] a taxonomy for binocular stereo dense algorithms is proposed. The authors decompose the stereo algorithms into four steps: matching cost computation, cost (support) aggregation, disparity computation/optimization and disparity refinement.

Significant improvements have been proposed over the last years. The most efficient algorithms tackle the correspondence problem by using disparity optimization methods. The aim is to enforce the smoothness assumption on both vertical and horizontal axes. Among the recently proposed optimization techniques, graph cut and belief propagation [7] seem to produce interesting results. Existing algorithms[8], [9], [10], [11], [12], [13], [14], [15] are computationally expensive and resource demanding. Other techniques such as scanline optimization [6], [16] and dynamic programming [6], [17], [18] provide accurate results with reasonable performance [6], [16], [17], [18], [19], [20]. Other reliable techniques use projectors as programmable light sources for active vision techniques using structured light range finding [8],

[9], [20], [21], [22], [23], photometry-based reconstruction [19], [24], relighting [25], light transport analysis [16], [26] and depth from defocus [27]. The main advantage of projecting a set of coloured patterns onto a scene is that it eases the problem of correspondences [21], but the method is inadequate for real-time processing. An evaluation of several algorithms can be found in [6]. Since the seminal work of creating multiple camera networks [28], tele-immersion became an important element for the next generation of live and interactive 3DTV applications. The goal of these techniques is to allow people at different physical locations to share a virtual environment.

1.3. Neuromorphic Silicon Retina

Biological retinas encode visual information differently than classical frame-based cameras. Transmitting only information on parts of the scene that have been changing (e.g. luminosity), biological retinas avoid acquiring and transmitting redundant data while adding precise time information. In the late 1980s, the first neuromorphic vision sensor mimicking the various behaviours of the biological retina was proposed by Mahowald [29]. It introduced an analog photoreceptor that transforms the perceived light intensity into an output voltage following a logarithmic mapping. Delbruck and Mead improved the design by adding active adaptation [30] and Kramer further added polarity encoding luminosity intensity change [31]. A review of some of the history and recent developments in artificial retina sensors can be found in [32].

The experiments reported in this paper were conducted using the dynamic vision sensor (DVS) described in [33]. The DVS is a 128×128 pixel resolution Address Event Representation (AER) silicon retina sensor that asynchronously generates response to relative light intensity variations. Pixels operate autonomously and encode temporal contrast, i.e. log intensity changes of a programmable magnitude, into events carrying the active pixel's array address and polarity of change (ON/OFF). The output channel is a parallel, continuous-time, digital bus that asynchronously transmits the Address Events. The data volume of such a self-timed, event-driven sensor depends essentially on the dynamic contents of the target scene as pixels that are not visually stimulated do not produce output. Due to the pixel-autonomous, asynchronous operation, the temporal resolution is not limited by an externally imposed frame rate. However, the asynchronous stream of events carries only change information and does not contain absolute intensity information; there are no conventional image data in the sense of gray-levels.

1.4. Stereo-correspondence in neuromorphic engineering

Mahowald et al. [34] implemented cooperative stereovision in a neuromorphic chip in 1989. The resulting sensor was composed of two 1D pixel arrays of 5 neuromorphic pixels each. The use of local inhibition driven along the line of sight implemented the uniqueness constraint (one pixel from one view is associated to only one in the other, except during occlusions), while the lateral excitatory connectivity gave more weight to coplanar solutions to discriminate false matches from correct ones. This method requires a great amount of correlator units to deal with higher resolution sensors.

In 2008, Shimonomura, Kushima and Yagi implemented the biologically inspired disparity energy model to perform stereovision with two silicon retinas [35]. They simulated elongated receptive fields to extract the disparity of the scene and control the vergence of the cameras. The approach is frame-based and allows to extract coarse disparity measurements to track object in 3D.

Kogler et al. [36] have described a frame-based usage of the event-based DVS cameras in 2009. They designed an event-to-frame converter to reconstruct event frames and then tested two conventionalk stereo vision algorithms, one window-based and one feature-based using center-segment features [37].

Delbruck has implemented a real event-based stereo tracker that tracks the position of a moving object in both views using an event-based median tracker and then reconstructs the position of the object in 3D [38]. This efficient and fast method lacks resolution on smaller features and is sensitive to noise when too many objects are present.

In 2011, Rogister et al. [39] proposed an asynchronous event-based binocular stereo matching algorithm combining epipolar geometry and timing information. Taking advantage of the high temporal resolution and the epipolar geometry constraint they provided a truely event-based approach for real-time stereo matching. However, this method is more sensitive to noise than

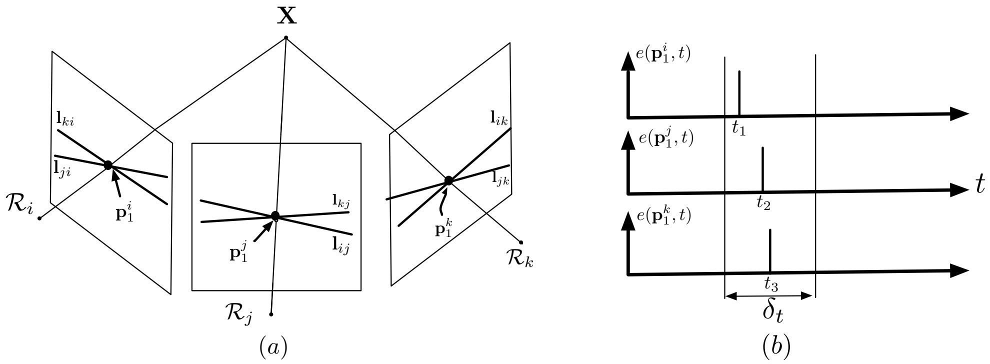

Figure 1: (a) Epipolar planes and lines illustrated for three cameras. A 3D point $\mathbf{X}$ is projected onto the three focal planes in $\mathbf{p}_1^i$, $\mathbf{p}_1^j$ and $\mathbf{p}_1^k$. Each of them is at the intersection of two epipolar lines defined by the geometric configuration. (b) Events generated by $\mathbf{X}$ in each camera at time t are usually not recorded with the same date t, but rather different timestamps t_1, t_2, etc. due to a finite precision in synchronizing the cameras.

the one presented in this paper, mainly because it has a less constrained formulation: in the new method presented here, at least three instead of two cameras are required to triangulate 3D points.

2. Asynchronous N-Ocular Stereo Vision

2.1. Trinocular geometry

Adding more cameras in stereovision applications is a natural technique for solving the depth recovery problem. Additional sensors not only reduce the occurence of occlusions but also reinforces the epipolar constraint linking pairs of cameras. If the number of cameras is sufficient, the geometric constraint alone can be used to uniquely define a set of points projected onto each camera.

Figure 1 depicts the typical geometric configuration for a set of three cameras. A 3D point seen by the cameras R_i, R_j, R_k is projected onto their respective focal planes in $\mathbf{p}_1^i, \mathbf{p}_1^j, \mathbf{p}_1^k$. If $\mathbf{p}_1^i$ is fixed, then the epipolar constraint states that $\mathbf{p}_1^j$ (respectively $\mathbf{p}_1^k$) lies on an epipolar line in R_j (respectively in R_k). Technical details on the epipolar properties can be found in [40].

The same property is true if we consider $\mathbf{p}_1^j$ or $\mathbf{p}_1^k$ as fixed. Thus, $\mathbf{p}_1^i, \mathbf{p}_1^j, \mathbf{p}_1^k$ are uniquely defined as intersections of two epipolar lines on each focal plane. The unicity of the triplet is only true if the epipolar planes do not overlap.

The overlapping happens when all the focal points are coplanar or aligned (which is a special case of coplanarity). These degenerate cases can be reduced by adding more cameras.

The geometrical constraint can be expressed by a homogeneous scalar equation built from the following definitions:

- an event e occurring at time t, observed by the camera $\mathcal{R}_i$ at pixel $\mathbf{p}_u^i = (x, y)^T$ is a function taking value in $\{-1; 1\}$ (the subscript u indexes matched events across the sensor focal planes). Its value is equal to 1 when the contrast increases and -1 when it decreases. The event is therefore defined as $e(\mathbf{p}_u^i, t)$.
- a 3D point $\mathbf{X}$ generating events $e(\mathbf{p}_1^i, t)$, $e(\mathbf{p}_1^j, t)$ and $e(\mathbf{p}_1^k, t)$, is projected respectively as $\mathbf{p}_1^i$, $\mathbf{p}_1^j$ and $\mathbf{p}_1^k$ according to the relation :

$$\begin{pmatrix} \mathbf{p}_1^u \\ 1 \end{pmatrix} = P_u \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}, u \in \{i, j, k\}. \tag{1}$$

  $P_u = K_u \begin{pmatrix} R_u & -\mathbf{C}_u \end{pmatrix}$ is the projection matrix of $\mathcal{R}_u$. R_u and $\mathbf{C}_u$ are the extrinsic parameters and K the intrinsic ones (for more details on the projection matrix, the reader can refer to [40]).

  The image point $\mathbf{p}_1^i$ then satisfies the epipolar constraint:

$$(\mathbf{p}_1^j \ 1)^T F_{ij} \begin{pmatrix} \mathbf{p}_1^i \\ 1 \end{pmatrix} = 0, \tag{2}$$

F_{ij} is the fundamental matrix establishing the geometric relationship linking $\mathcal{R}_i$ to $\mathcal{R}_j$. $F_{ij} \begin{pmatrix} \mathbf{p}_1^i \\ 1 \end{pmatrix} = \mathbf{l}_{ij}(\mathbf{p}_1^i)$ is the epipolar line on $\mathcal{R}_j$, associated to $\mathbf{p}_1^i$. $\mathbf{p}_1^j$ belongs to $\mathbf{l}_{ij}(\mathbf{p}_1^i)$.
Threough similarl reasoning, all epipolar lines shown in figure 1 can be defined. If $\mathbf{p}_1^i$ and $\mathbf{p}_1^j$ are known, and the cameras are calibrated, then the $\mathbf{p}_1^k$ can be calculated as the intersection of the appropriate epipolar lines.

2.2. Trinocular spatio-temporal match

Estimating 3D from the cameras requires matching each triplet $\{\mathbf{p}_1^i, \mathbf{p}_1^j, \mathbf{p}_1^k\}$ produced by $\mathbf{X}$ at time t. Since the silicon retina sensors do not provide intensity information, only the geometric property presented in the previous section can be used in conjunction with the highly accurate timing of the events. Let us define the set of events occurring within a time window around time t :

$$S^i(t) = \left\{ e(\mathbf{p}^i, t') | \mathbf{p}^i \in \mathbb{R}^2 \text{ and } t, t' \in \mathbb{R}^+, |t' - t| < \frac{\delta_t}{2} \right\}. \tag{3}$$

S^i defines a temporal neighbourhood of events captured by $\mathcal{R}_i$ that occur around t. Such sets are defined for each camera. Because of non-perfect synchronization of the cameras, it is unlikely that matched events are timestamped with the same t (see figure 1).

In a similar way, we define the set of events geometrically close to $\mathbf{l}_{ij}(\mathbf{p}^i)$:

$$M^j(e(\mathbf{p}^i, t)) = \left\{ e(\mathbf{p}^j, t') \in S^j(t) | d(\mathbf{p}^j, \mathbf{l}_{ij}(\mathbf{p}^i)) < \Delta_p \right\}, \tag{4}$$

where $d(\mathbf{p}^j, \mathbf{l}_{ij})$ is the euclidean distance of $\mathbf{p}^j$ to $\mathbf{l}_{ij}$. The image points $\mathbf{p}_1^j, \mathbf{p}_1^k$, elements of sets $M^j(\mathbf{p}_1^i, t)$ and $M^k(\mathbf{p}_1^i, t)$ respectively, are matched to $\mathbf{p}_1^i$ if they minimize both $|t - t'|$ and $d(\mathbf{p}_1^i, \mathbf{l}_{ij})$ defined in Eq. (3) and (4).
Due to the finite precision of the visual acquisition in space and time, the matching process is prone to produce erroneous matches because of additional ambiguities beside the ones induced by degenerate cases. The motivation to use more than just two cameras is also given by [41]. The authors show that the use of a third camera reduces the number of ambiguities by a factor of 10 when only geometric constraints can be used. For event-based sensors, the accurate timing adds decisive complementary constraints.

Based on the previous definitions, we design the general trinocular point matching algorithm using temporal and spatial constraint as shown in algorithm 1. This matching algorithm requires a calibrated camera setup. Appropriate calibration can be achieved with the techniques presented in [42] if only the fundamental is needed, or the one from [43] if the projection matrix is also required. The algorithm can be extended to n cameras with minimal changes.

2.3. Stereo match selection using bayesian inference

A triplet of matched events $m_n = \{e(\mathbf{p}_n^i, t), e(\mathbf{p}_n^j, t), e(\mathbf{p}_n^k, t)\}$ is a true match if the events are generated in response to a same stimulus, at the same

Algorithm 1 Trinocular event-based stereo matching algorithm

```
Require: Three cameras $\mathcal{R}_i$, $\mathcal{R}_j$, $\mathcal{R}_k$
Require: $F_{ij}$, $F_{ik}$, $F_{jk}$, estimations of the fundamental matrix for each pair
    of cameras
for all events $e(\mathbf{p}_n^i, t)$ in sensor $\mathcal{R}_i$ do
    Determine the set of events $S^j(t)$ from sensor $\mathcal{R}_j$
    Determine the set of events $S^k(t)$ from sensor $\mathcal{R}_k$
    Compute the epipolar line $\mathbf{l}_{ij} = F_{ij}\binom{\mathbf{p}_n^i}{1}$
    Compute the epipolar line $\mathbf{l}_{ik} = F_{ik}\binom{\mathbf{p}_n^i}{1}$
    Determine the subset of possible matches $M^j(e(\mathbf{p}_1^i, t)) \subset S^j(t)$
    for all events $e(\mathbf{p}_k^j, t) \in M^j(e(\mathbf{p}_1^i, t))$ do
        Compute the epipolar line $\mathbf{l}_{jk} = F_{jk}\binom{\mathbf{p}_k^j}{1}$
        Compute intersection between $\mathbf{l}_{jk}$ and $\mathbf{l}_{ik}$
        if $e(\mathbf{p}_n^i, t) \in S^i(t)$, $e(\mathbf{p}_n^j, t) \in S^j(t)$, $e(\mathbf{p}_n^k, t) \in S^k(t)$ complies to the
        trinocular constraint then
            Create match $m_n = \{e(\mathbf{p}_n^i, t), e(\mathbf{p}_n^j, t), e(\mathbf{p}_n^k, t)\}$ and add it to the
            list of found matches $T(\mathcal{R}_i, \mathcal{R}_j, \mathcal{R}_k)$
        end if
    end for
end for
return $T(\mathcal{R}_i, \mathcal{R}_j, \mathcal{R}_k)$
```

time. The triplet is mismatched otherwise. For each m_n, a corresponding 3D point $\hat{\mathbf{X}}_\mathbf{n} = (x, y, z)^T$ can be estimated as the intersection of the back-projected rays by "inverting" Eq. 1:

$$\hat{\mathbf{X}}_\mathbf{n} = \bigcap_{u \in \{i,j,k\}} \lambda_u R_u^{-1} K_u^{-1} \begin{pmatrix} \mathbf{p}_n \\ 1 \end{pmatrix} + \mathbf{C}_u, \tag{5}$$

where λ_u is a scalar.

If a given match m_1 is actually a wrong match, then $e(\mathbf{p}_1^i, t_1)$, $e(\mathbf{p}_1^j, t_2)$ and $e(\mathbf{p}_1^k, t_4)$ are not events induced by the same stimulus in the scene. The set m_1 yields a 3D point which either does not physically exist at time t, or at the location of $\hat{\mathbf{X}}_\mathbf{n}$ in the real scene.

The probability for a set $m_1 = \{\mathbf{p}_1^i, \mathbf{p}_1^j, \mathbf{p}_1^k\}$ at time t, to be a correct match is related to the spatio-temporal neighbourhood of $\hat{\mathbf{X}}_\mathbf{1}$. Because scenes are usually composed by geometric structures which generate edges in the sensors' focal planes, it is unlikely that an isolated 3D point $\hat{\mathbf{X}}_{\mathbf{1}}$ exists in the scene. We add a statistical constraint using bayesian inference to sort outliers from correct matches. We first define the set of potential matches contained in a spatio-temporal neighbourhood of m_1:

$$W(m_1) = \left\{ m_n \in T(\mathcal{R}_i, \mathcal{R}_j, \mathcal{R}_k) | d_s(m_1, m_n) \leq \delta_s, \bar{d}_t(m_1, m_n) \leq \delta_t \right\}, \quad (6)$$

with

- $d_s(m_1, m_2) = ||\hat{\mathbf{X}}_{\mathbf{1}} - \hat{\mathbf{X}}_{\mathbf{2}}||$,
- $\bar{d}_t(m_1, m_2)$, the mean duration between the 6 events defined in m_1 and m_2.

δ_s and δ_t are the spatial and the temporal radii of the neighbourhood (see figure 2). The two components are decoupled to allow a fine adjustment of the neighbourhood.

Given $W(m_1)$, the probability of m_1 being a correct match is deduced from Bayes' rule :

$$P(m_1 | W(m_1)) = \frac{P(W(m_1)|m_1)P(m_1)}{P(W(m_1))}. \quad (7)$$

2.3.1. Prior

Prior probability is established from the matching algorithm presented in section 2.1. The reliability of each match m_n is defined according to how well they comply with the spatio-temporal constraint i.e. how far temporally and spatially the events are from the epipolar intersections given a time t. Typically a gaussian distribution is fitted on the matching results.

2.3.2. Likelihood

The Likelihood of having a correctly matched triplet m_n is assumed to increase inversely with its distance to a triplet of matched events that is labelled as correct. Following this hypothesis, the conditional probability of m_n according to its spatio-temporal neighbour m_1 is defined as:

$$P(m_n = 1 | m_1) = \begin{cases} \mathcal{N}(0, \sum) & \text{if } m_1 = 1 \\ k & \text{if } m_1 = 0 \end{cases}, \quad (8)$$

where $\mathcal{N}(\mu, \sum)$ is a bivariate gaussian distribution of mean value μ and covariance matrix $\sum$. The probability of having a correct match when its

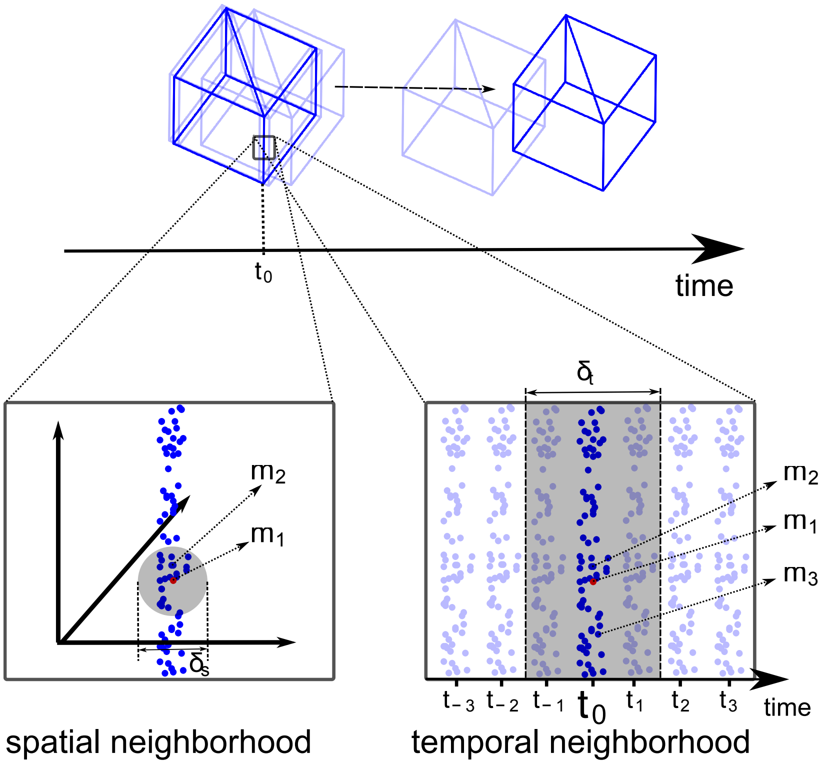

Figure 2: Probability of matches in the spatio-temporal neighbourhood of m_1 limited by δ_t and δ_s. m_2 is in the spatio-temporal neighbourhood of m_1 therefore has high probability of being correct. m_3 is in the temporal neighbourhood of m_1 but outside the spatial neighbourhood being therefore probably an incorrect match.

neighbour is not correct ($P(m_n = 1|m_1 = 0)$) is usually small, as isolated 3D points are unlikely to exist in real scenes. k is established based on observations from experimental results.

If we assume that the probability for a given $m_n \in W(m_1)$ to be a correct match, depends only on m_1 (i.e. 2 triplets of events m_i, m_j in $W(m_1)$ are independent), then the joint probability $P(W(m_1)|m_1)$ is given by:

$$P(W(m_1)|m_1) = \prod_{m_n \in W(m_1)} P(m_n|m_1), \tag{9}$$

2.3.3. Posterior

The posterior $P(m_1|W(m_1))$ is computed continuously over time according to Eq. (7) in order to update the 3D reconstruction model. The 3D structure is therefore progressively reconstructed. During initial stages few sparse matches are observed and the model is poor. As more matches are found in further iterations, matches belonging to edges are given higher probability and the model is progressively refined.

2.4. Synchronization

The spatio-temporal matching requires the accurate synchronization of all cameras since matched events result from a common stimulus at time t. The synchronization is achieved using an external trigger signal. However the synchronization accuracy is limited due to several factors:

- non-isotropic stimuli or non identical pixel sensitivities induce different event recording times,
- varying transmission latencies of the sensor output buses due to event collisions. When multiple photoreceptors fire at the same time the sensor's bus arbiters serialize event output, thus delaying the real occurrence and potentially shuffling the firing order of events.

This synchronization uncertainty is referred to as event jitter and can be measured experimentally. We have placed an LED blinking at 10Hz in front of the 6-cameras system. The measured response is shown in figure 3: all

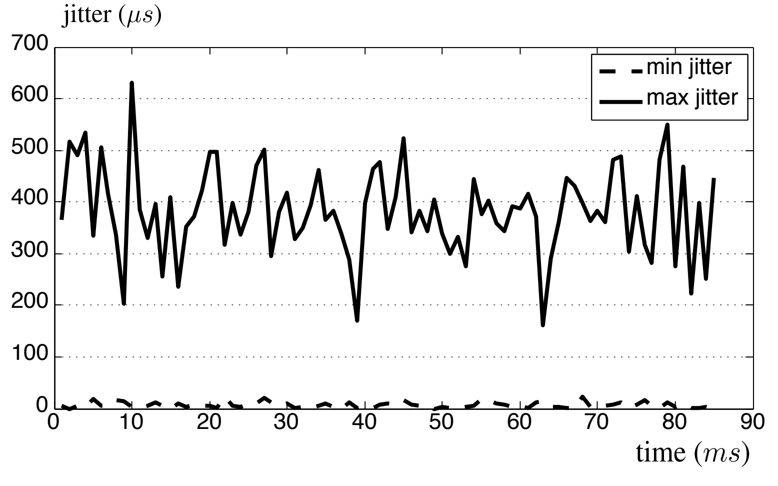

Figure 3: Jitter between cameras' response to a blinking led. The curves shows the minimum and maximum jitter between all possible combination of two cameras.

cameras responded within a maximum delay of $631\mu s$ throughout the experiment. In average all cameras responded within a $382\mu s$ window.

The variable relative delays between sensors limit the time accuracy at which events are matched. Moreover, the delays are also scene dependent, making the task even more difficult. To achieve correct spatio-temporal matches, prior assumptions about the scene should be made in order to establidh the upper bound of the timing accuracy.

2.5. N-ocular stereo matching

The trinocular configuration provides the minimal geometric constraint to uniquely identify the set of matched events. The matching algorithm presented in section 2.1 can however be extended to more cameras. Two variations are presented showing different advantages:

- Each camera contributes to enforce the epipolar constraint and the time consistency: matched points at time t are on intersections of a set of epipolar lines. The reliability of matched points increases with the number of used cameras.
- Events are matched by grouping exhaustively all subset of three cameras. For N cameras, $\binom{N}{3}$ unique trinocular configurations exist.

The number of matched events using the first variant decreases with the number of cameras as increasing this number increases temporal and geometrical constraints. Resulting 3D reconstructions often contain too little succesfully matched events and are not sufficient to provide complete representations of 3D shapes. In addition, the computational effort increases drastically with the number of cameras.

The second variant delivers more matched events resulting in a denser reconstruction, however including more false matches. Considering these observations, the best strategy for the event-based 3D reconstruction to be dense and fast enough is to combine the second variant with applying bayesian inference selsction.

3. Experimental results

A setup of six cameras has been used to evaluate the spatio-temporal 3D reconstructions principle. The six DVS cameras are synchronized using an external clock. The sensors are also geometrically calibrated using the method given in [43]. The achieved calibration accuracy is sub-pixelic.

Examples of reconstructions are shown in figure 4 for three objects moving in front of the cameras. The time windows used for matching the events are defined in accordance with the jitter problem presented in section 2.4: $500\mu s$ is used for a swinging cube, $1000\mu s$ for a waving hand and $2000\mu s$ for the a moving human face.

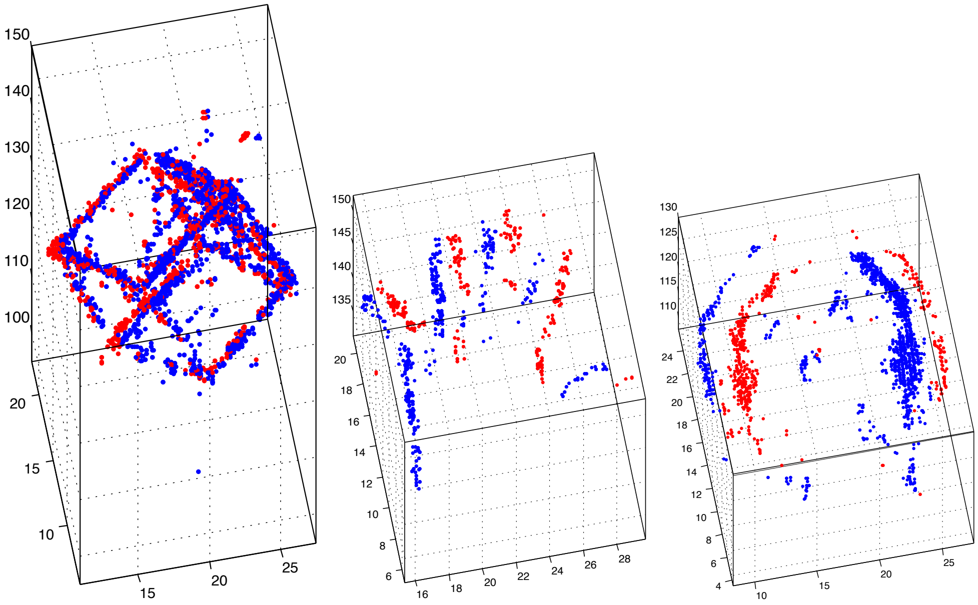

Figure 4: Example of reconstructions obtained using the event-based trinocular algorithm presented in section 2.5: (left) wireframe cube, (center) hand , (right) human face. The colors encode the polarity of the events producing the reconstructed 3D points. They give a hint to the motion's direction.

3.1. Reconstruction Evaluation

The 3D reconstructions are evaluated in order to quantify the algorithm's quality. Two techniques are proposed to measure reconstruction errors:

- if the ground truth is available, we measure the differences between the reconstructed shapes and the original,

- if the ground truth does not exist, we project the reconstructed shape onto the cameras that were not part of the actual triplet used for reconstruction.

For the wireframe cube, the geometric ground truth is perfectly known and is compared with the reconstructions at each new incoming event. The ground truth's 3D points are first fitted to the reconstructed points using a 3D points set registration algorithm (ICP) [44]. Then the mean distance, normalized by the edge length c, of all the reconstructed points to the ground truth is computed:

$$\epsilon = \frac{\sum_{i=1}^{n} e_i^2}{n.c^2}. \tag{10}$$

(see figure 5 for an illustration of distances e_i of reconstructed points (circles) to the ground truth model (plain curve)).

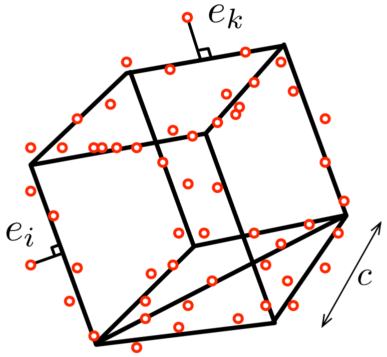

Figure 5: Distance of reconstructed points to the ground truth model.

Figure 6 summarizes the normalized reconstruction errors of the moving wireframe cube using all camera combination triplets out of a set of 3 to 6 cameras. The error is computed for the entire sequence with and without Bayesian inference. One can see that increasing the number of sensors also increases the reconstructed 3D points. Furthermore using Bayes' rule to filter erroneous matches The relative mean error gives an idea of how reliable the reconstructed shape is. In this case, the mean error is around 2.5% for a set of 5 sensors using the trinocular algorithm alone. The same error is reduced to 1.5% if the Bayesian inference is applied. For a 6 cameras system, these values are reduced to 1.5% and 0.5% respectively. The reconstruction errors

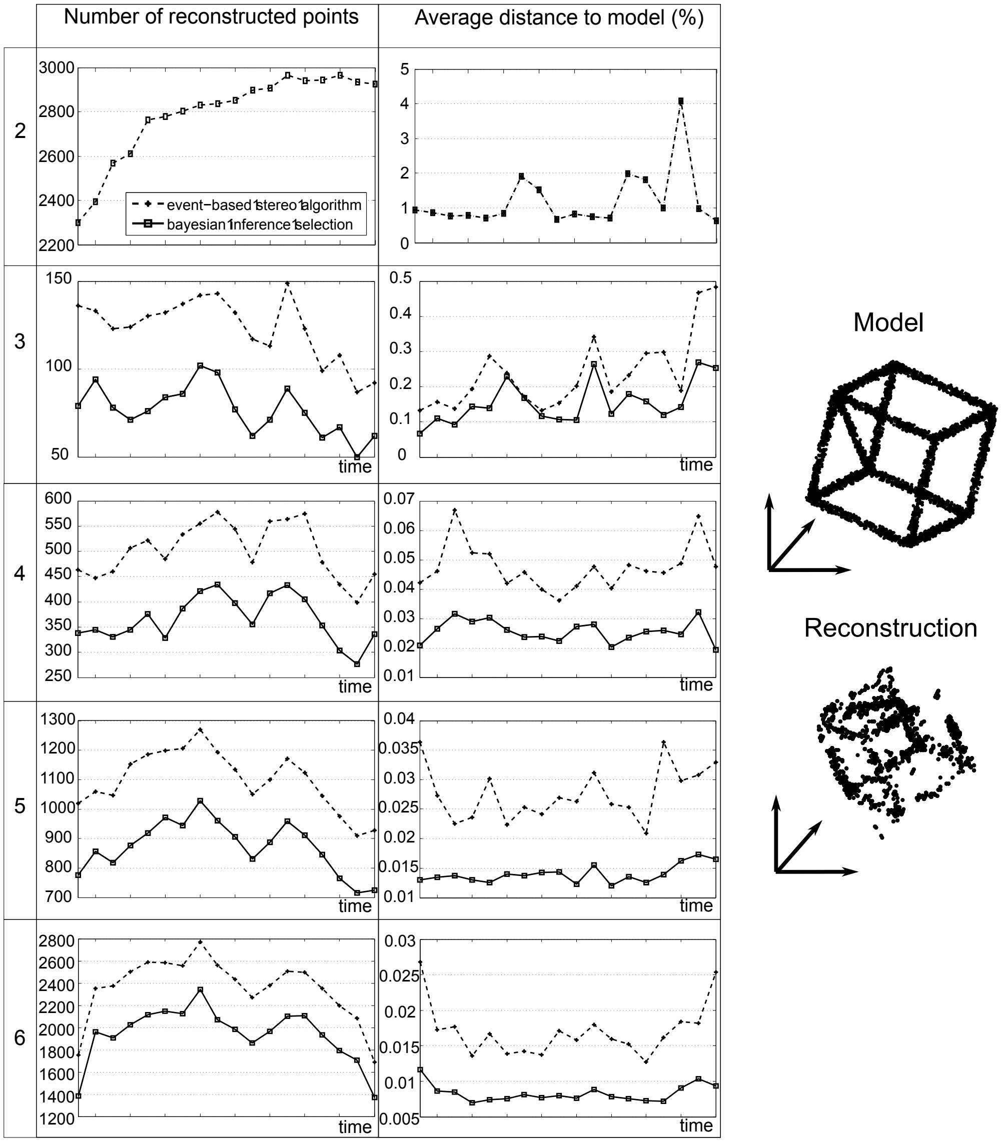

Figure 6: Reconstruction errors of the wireframe cube: the first column curves show the number of reconstructed points while in the second column, curves are showing reconstruction's errors. For comparison purpose, we put in the top row, the results achieved by the method explained in [39]. Rows 2 to 5 are the results produced by the method we have introduced in this paper, with the number of cameras increasing from 3 to 6. Reconstructions quality is also measured with (dashed curves) and without (plain curves) Bayesian inference.

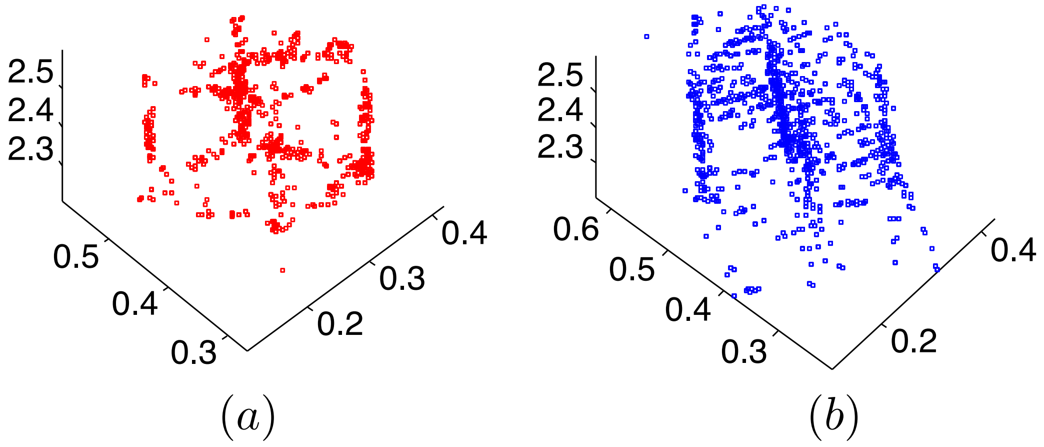

Figure 7: Reconstructed cube using the trinocular constrained algorithm (a) and using the event-based stereovision algorithm for 2 cameras (b). In the second case, the reconstruction result is visibly less accurate.

for 2 sensors using the method from [39] are also plotted in order to show the performance of the trinocular algorithm since relative errors never exceed 1% while with the technique given in [39] can reach 400%. 3D reconstruction results shown in figure 7 give a quick visual assessement of the reconstructions performance: the reconstructed points of the cube (b) is more scattered than the ones using 3 or more cameras.

With regard to this preliminary result, we state that the wireframe cube is usually reconstructed with acceptable accuracy. Any other quality assessment giving scores similar to the cube reconstruction are therefore assumed to correspond to reconstructions of sufficient quality.

Reconstructions for which the ground truth is unavailable require another technique for evaluating their accuracy. We apply a variation of the method presented in [45]. Assuming that a set of the 3D object is built from three cameras $\mathcal{R}_j, \mathcal{R}_k, \mathcal{R}_l$ and given another camera $\mathcal{R}_i$ with $i \notin \{j, k, l\}$, the reconstructed objects are evaluated as follows:

- the objects are projected onto $\mathcal{R}_i$. This operation gives rise to a frame.

- a frame is also built by integrating the events captured by $\mathcal{R}_i$ over the time window defined for the matching algorithm (e.g. $500\mu s$ for the cube).

- the ratio of pixel differences given by subtracting both frame produces the projection error.

Figures (8, 9, 10) show the reconstruction errors for the three sequences.
The estimated error is low for the cube on all cameras: around 3% of error for each sensor. Since the cube reconstruction has already been shown to be accurate, an error of this magnitude is considered as a good indicator of a reliable reconstruction. For both the hand and the face sequences, the estimated errors have the same order of magnitude (3% for the hand and 5% for the face). We can therefore deduce that the trinocular algorithm is providing sufficiently accurate 3D reconstructions.

4. Processing time

The processing time is a critical issue especially for real-time applications. Computational effort obviously increases with the number of cameras used. In figure 11 the processing time with respect to the number of reconstructed points for the three sequences are shown for sets of 3 to 6 cameras. A remarkable observation is the linearity of the processing time for whatever number of cameras illustrating the linear complexity of the global reconstruction process.

5. Discussion

The Event-based matching approach shows the possibility to recover 3D from time and geometric consideration only. Two variants of the matching algorithm are tested for 3D reconstructions: the first method uses all possible combination of three sensors to compute 3D points from the events while the second method uses all sensors to enforce the epipolar constraint.
The first method gives the best compromise between the reconstruction accuracy, density and computation time. Since the reconstruction complexity is likely linear as suggested in section4, we expect the algorithm being largely optimizable. The algorithm's runtime is large on non compiled programming

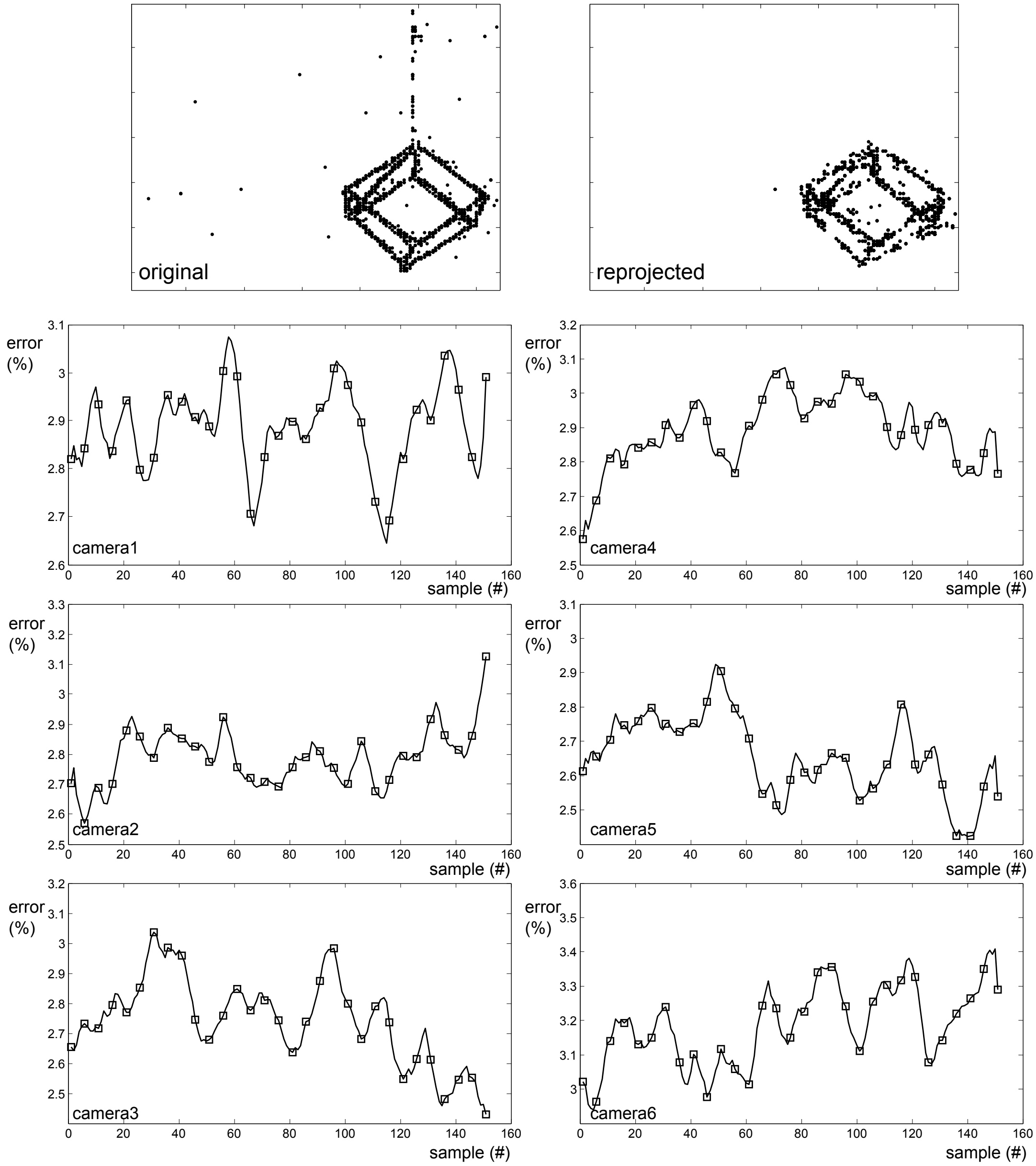

Figure 8: Reprojection errors on each of the 6 cameras. For each camera $\mathcal{R}_i$ that is tested, 3D cubes built from any combination of 3 other cameras $\mathcal{R}_{j,k,l}$ are projected onto $\mathcal{R}_i$. The obtained frame is then compared to the frame built by integration. Mean projection errors are around 3%.

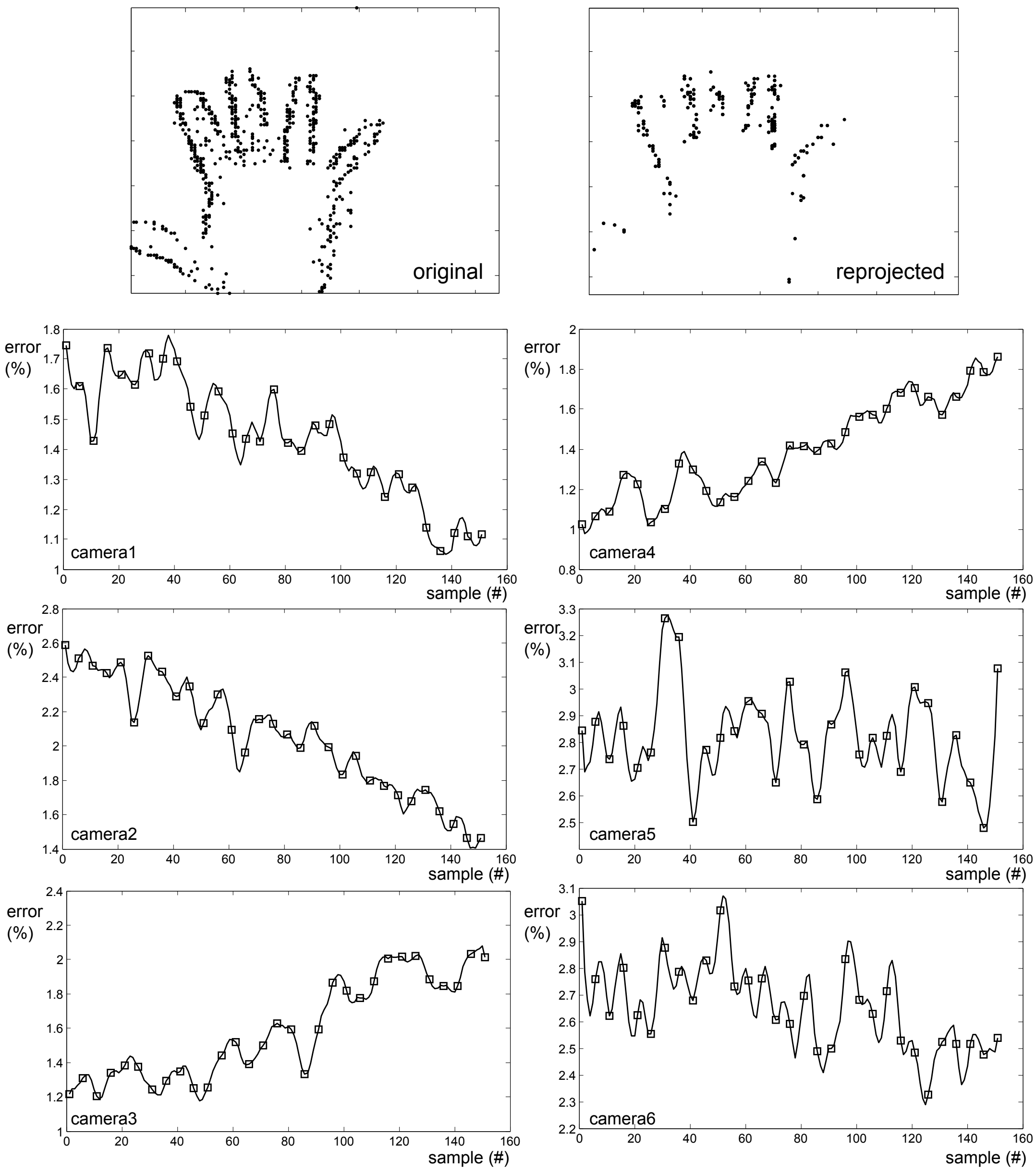

Figure 9: Reprojection errors estimated for the hand, using the same method as for the cube. Mean projection errors do not exceed 4%. For the first four cameras (left column and the top right curves), the error curves are showing constant increase/decrease. This is due to the hand leaving/entering the field of view of the cameras.

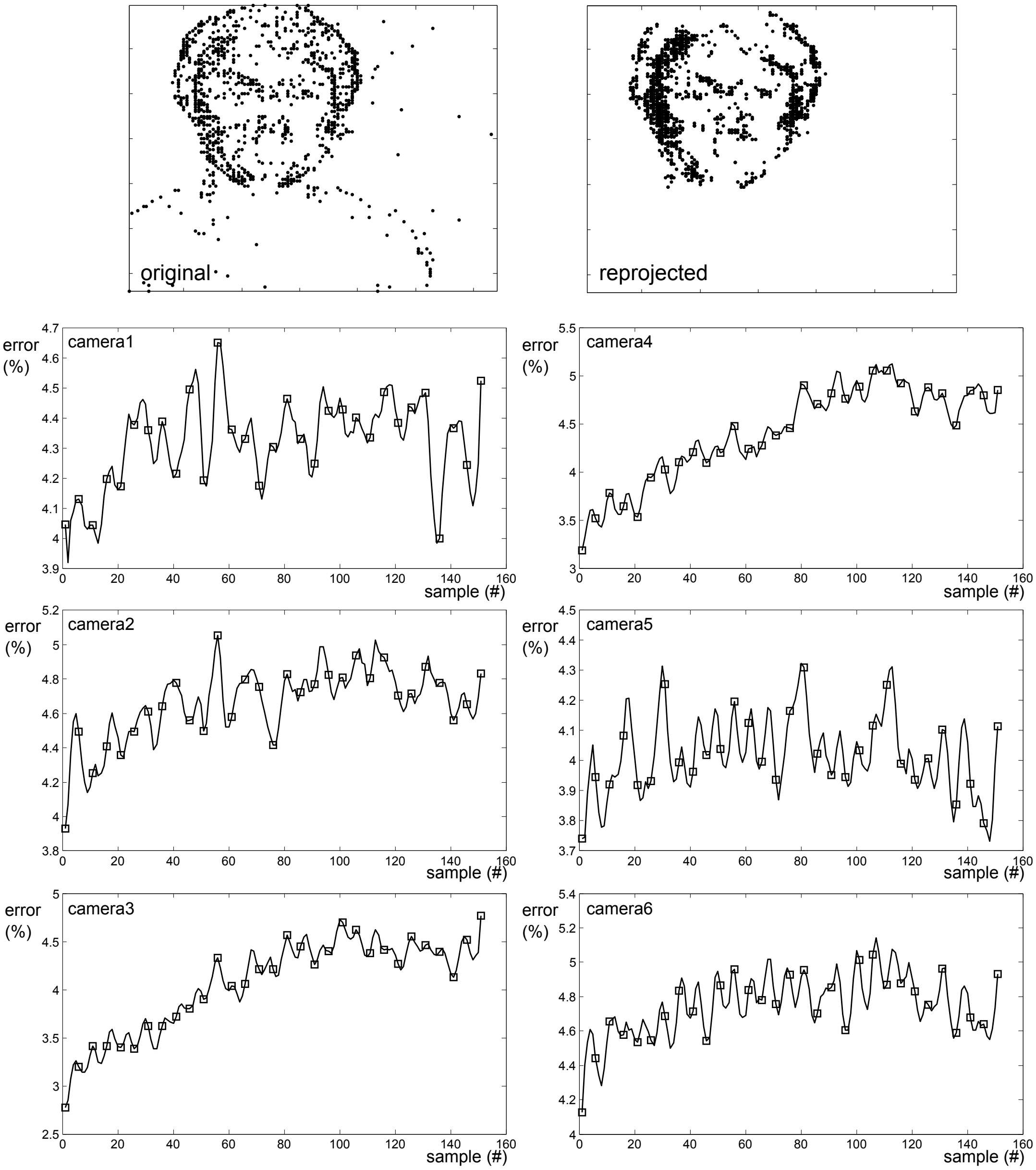

Figure 10: Difference between an image frame of events and reprojection of the reconstructed face on each of the 6 cameras. The mean projection errors in this sequence is not exceeding 5%.

languages such as Matlab, however it is very likely that processing time can

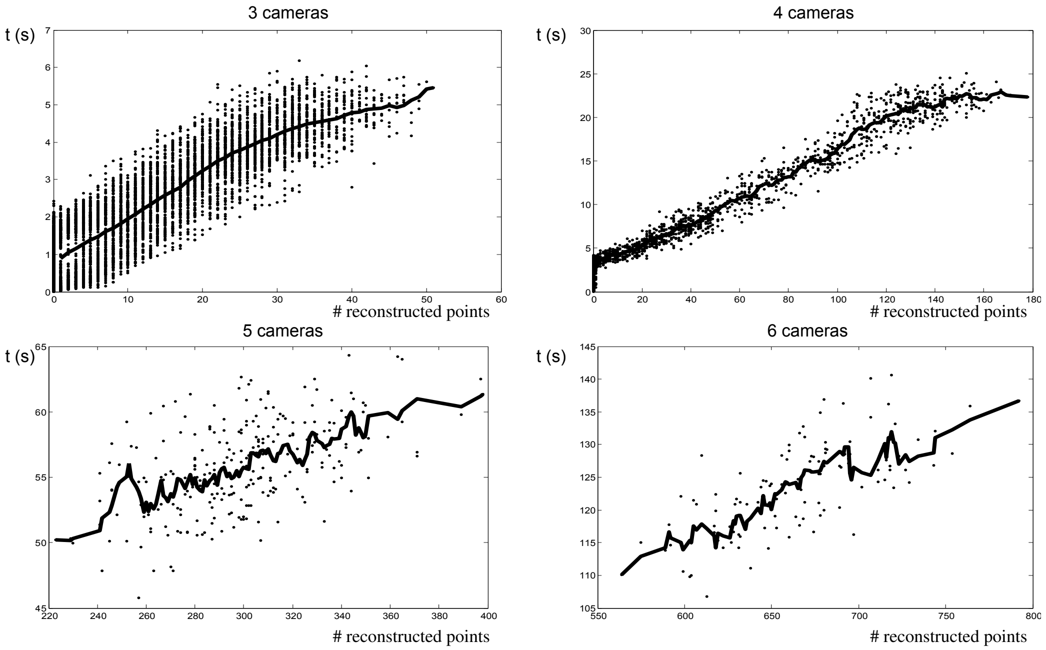

Figure 11: Processing time as function of the number of reconstructed points, for 3 to 6 cameras. The mean processing time is represented by the plain curves.

largely be reduced and meet real-time constraints when using a compiled programming language such as C.

6. Conclusion

This paper presents an asynchronous event-based stereo algorithm applied to multi-cameras systems of neuromorphic artificial retina sensors. The event-based stereovision opens a new perspective for recovering 3D information from sparse and asynchronous spatio-temporal signals. Because the temporal contrast DVS sensor does not capture intensity information, no complex characterization of pixels can be computed as it is done in traditional stereovision. We developed a new class of algorithms based only on geometrical and temporal constraints. This approach tackles the stereovision problem in a bio-inspired manner.

The achieved reconstructions are highly accurate despite the relatively low spatial resolution of the used sensor (128 × 128 pixels). The precise event timing provides a mean to overcome this limitation and substantiates that an event-based vision sensor is the ideal device to capture dynamic scenes. This work extends and pushes to a much higher level the neuromorphic stereovision formulation initiated in [39]. Here, events are processed individually and not as clustered sets, which then reduced the temporal accuracy of the individual events.

One major difficulty in establishing spatio-temporal stereovision is limited temporal precision of the visual information. Since time constitutes critical information, highly precise event timing is required by the algorithms, hence needs to be established at the sensor level and transmitted to the processing stage. Effective timing accuracy is limited by physical constraints of the sensor and system hardware such as e.g. CMOS device mismatch leading to intra-chip and inter-chip variations of contrast sensitivity and event latencies, event bus congestions, and timestamping quantization errors. However, it has been demonstrated that existing event-based sensor technology [33] is applicable in this context while any existing standard frame-based vision sensor with its temporal resolution limited by the frame rate would not be adequate.

Future work in this area is going to be based on new generations of event-based vision sensors that have become available recently. E.g. ATIS [46] features a five-fold increase in the number of pixels (QVGA) along with a corresponding improvement in spatial resolution while at the same time substantially reduces event latency and jitter, thus improving temporal resolution. Furthermore, in addition to temporal contrast events, the sensor outputs absolute pixel intensity information also encoded in the timing of asynchronous events, providing additional constraints for stereo event matching. Another recent DVS development [47] improves on the contrast sensitivity, allowing for the inclusion of more low-contrast visual information such as texture details. The future perspective of event-based stereo-vision is to push towards higher accuracies both in time and space, eventually aiming for real-time textured 3D reconstruction.

Funding

This research was funded by the french national Labex program ”Lifesenses”.

References

[1] F. Gonzalez, R. Perez, Neural mechanisms underlying stereoscopic vision, Progress in Neurobiology 55 (1998) 191 – 224.

[2] A. Bergua, W. Skrandies, An early antecedent to modern random dot stereograms -'the secret stereoscopic writing' of ramon y cajal., International Journal of Psychophysiology 36 (2000) 69–72.

[3] Parvati, Dev, Perception of depth surfaces in random-dot stereograms : a neural model, International Journal of Man-Machine Studies 7 (1975) 511 – 528.

[4] I. Jeremiah, Nelson, Globality and stereoscopic fusion in binocular vision, Journal of Theoretical Biology 49 (1975) 1 – 88.

[5] J. C. A. Read, A bayesian approach to the stereo correspondence problem, Neural Comput. 14 (2002) 1371–1392.

[6] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vision 47 (2002) 7–42.

[7] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, C. Rother, A comparative study of energy minimization methods for markov random fields with smoothness-based priors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 1068–1080.

[8] P. M. Will, K. S. Pennington, Grid coding: a preprocessing technique for robot and machine vision, in: Proceedings of the 2nd international joint conference on Artificial intelligence, IJCAI'71, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1971, pp. 66–70.

[9] J. Davis, D. Nehab, R. Ramamoorthi, S. Rusinkiewicz, Spacetime stereo: A unifying framework for depth from triangulation, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 296–302.

[10] A. Klaus, M. Sormann, K. Karner, Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, in: Proceedings of the 18th International Conference on Pattern Recognition - Volume 03, ICPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 15–18.

[11] Y. Taguchi, B. Wilburn, C. L. Zitnick, Stereo reconstruction with mixed pixels using adaptive over-segmentation, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (2008) 1–8.

[12] Z.-F. Wang, Z.-G. Zheng, A region based stereo matching algorithm using cooperative optimization, in: IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2008, pp. 1–8.

[13] Q. Yang, L. Wang, R. Yang, H. Stewénius, D. Nistér, Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling, IEEE Transation on Pattern Analysis and Machine Intelligence 31 (2009) 492–504.

[14] L. Xu, J. Jia, Stereo matching: An outlier confidence approach, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), IEEE on European Conference on Computer Vision., volume 5305 of *Lecture Notes in Computer Science*, Springer Berlin, Berlin, Heidelberg, 2008, pp. 775–787.

[15] Q. Yang, R. Yang, J. Davis, D. Nistér, Spatial-depth super resolution for range images., in: IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2007.

[16] S. K. Nayar, G. Krishnan, M. D. Grossberg, R. Raskar, Fast separation of direct and global components of a scene using high frequency illumination, in: ACM SIGGRAPH 2006 Papers, pp. 935–944.

[17] L. Wang, M. Liao, M. Gong, R. Yang, D. Nister, High-quality real-time stereo using adaptive cost aggregation and dynamic programming, in: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), pp. 798–805.

[18] O. Veksler, Stereo correspondence by dynamic programming on a tree, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2 of *CVPR '05*, IEEE Computer Society, Washington, DC, USA, 2005, pp. 384–390.

[19] T. E. Zickler, P. N. Belhumeur, D. J. Kriegman, Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction, Int. J. Comput. Vision 49 (2002) 215–227.

[20] M. Young, E. Beeson, J. Davis, S. Rusinkiewicz, R. Ramamoorthi, Viewpoint-coded structured light, in: IEEE Conference on Computer Vision and Pattern Recognition - CVPR 2007, pp. 1–8.

[21] L. Zhang, B. Curless, S. M. Seitz, Rapid shape acquisition using color structured light and multi-pass dynamic programming, in: First International Symposium on 3D Data Processing Visualization and Transmission - 2002, pp. 24–36.

[22] B. Curless, M. Levoy, Better optical triangulation through spacetime analysis, in: International Conference on Computer Vision, Stanford University, Stanford, CA, USA, 1995.

[23] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition, CVPR'03, IEEE Computer Society, Washington, DC, USA, 2003, pp. 195–202.

[24] A. Hertzmann, S. M. Seitz, Shape and materials by example: a photometric stereo approach, in: Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition, CVPR'03, IEEE Computer Society, Washington, DC, USA, 2003, pp. 533–540.

[25] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, P. Debevec, Performance relighting and reflectance transformation with time-multiplexed illumination, in: ACM SIGGRAPH 2005 Papers, pp. 756–764.

[26] P. Sen, B. Chen, G. Garg, S. R. Marschner, M. Horowitz, M. Levoy, H. P. A. Lensch, Dual photography, ACM Trans. Graph. 24 (2005) 745–755.

[27] L. Zhang, S. Nayar, Projection defocus analysis for scene capture and image display, in: ACM SIGGRAPH 2006 Papers, pp. 907–915.

[28] T. Kanade, H. Saito, S. Vedula, The 3D room: digitizing time-varying 3D events by synchronized multiple video streams, Technical Report, CMU, 1998.

[29] M. Mahowald, VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function, Ph.D. thesis, California Institut of Technology, 1992.

[30] T. Delbrück, C. A. Mead, Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits, in: C. Koch, H. Li (Eds.), Vision Chips: Implementing vision algorithms with analog VLSI circuits, IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 139–161.

[31] J. Kramer, An on/off transient imager with event-driven, asynchronous read-out, in: in Proc. IEEE International Symposium on Circuits and Systems - ISCAS 2002, pp. 165–168.

[32] T. Delbrück, B. Linares-Barranco, E. Culurciello, C. Posch, Activity-driven, event-based vision sensors., in: ISCAS, IEEE, 2010, pp. 2426–2429.

[33] P. Lichtsteiner, C. Posch, T. Delbruck, A 128x128 120 db 15 us latency asynchronous temporal contrast vision sensor, IEEE Journal of Solid State Circuits 43 (2008) 566–576.

[34] M. Mahowald, T. Delbrück, Cooperative stereo matching using static and dynamic image features, in: C. Mead, M. Ismail (Eds.), Analog VLSI Implementation of Neural Systems, Kluwer Academic Publishers, 1989, pp. 213–238.

[35] K. Shimonomura, T. Kushima, T. Yagi, Binocular robot vision emulating disparity computation in the primary visual cortex, Neural Networks 21 (2008) 331 – 340.

[36] J. Kogler, C. Sulzbachner, W. Kubinger, Bio-inspired stereo vision system with silicon retina imagers, Computer Vision Systems (2009) 174–183.

[37] J. Shi, C. Tomasi, Good features to track, in: IEEE Conference on Computer Vision and Pattern Recognition, Cornell University, Ithaca, NY, USA, 1993.

[38] J. Lee, T. Delbruck, P. Park, M. Pfeiffer, C. Shin, H. Ryu, B. Kang, Gesture-based remote control using stereo pair of dynamic vision sensors, in: IEEE International Symposium on Circuits and Systems - ISCAS 2012.

[39] P. Rogister, R. Benosman, S. Ieng, P. Lichtsteiner, T. Delbruck, Asynchronous event-based binocular stereo matching, IEEE Transactions on Neural Networks 23 (2011) 347–353.

[40] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[41] H.-G. Maas, Complexity analysis for the establishment of image correspondences of dense spatial target fields, International Archives of Photogrammetry and Remote Sensing 29 (1992) 102–107.

[42] R. Benosman, S. Ieng, P. Rogister, C. Posch, Asynchronous event-based hebbian epipolar geometry, IEEE Transaction on Neural Network 22 (2011) 1723–1734.

[43] T. Svoboda, D. Martinec, T. Pajdla, A convenient multi-camera self-calibration for virtual environments, PRESENCE: Teleoperators and Virtual Environments 14 (2005) 407–422.

[44] D. Chetverikov, D. Stepanov, P. Krsek, Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm, Image and Vision Computing 23 (2005) 299–309.

[45] S. Sinha, M. Pollefeys, L. Mcmillan, Camera network calibration from dynamic silhouettes, in: IEEE conference on Computer Vision and Pattern Recognition, pp. 195–202.

[46] C. Posch, D. Matolin, 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.

[47] T. Serrano-Gotarredona, B. Linares-Barranco, A 128 x 128 1.5latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers, IEEE Journal of Solid-State Circuits 48 (2013).