On event-based motion detection and integration

Stephan Tschechne* Tobias Brosch Roman Sailer

Nora von Egloffstein Luma Issa Abdul-Kreem Heiko Neumann

Institute for Neural Information Processing
Ulm University, 89081 Ulm, Germany

ABSTRACT

Event-based vision sensors sample individual pixels at a much higher temporal resolution and provide a representation of the visual input available in their receptive fields that is temporally independent of neighboring pixels. The information available on pixel level for subsequent processing stages is reduced to representations of changes in the local intensity function. In this paper we present theoretical implications of this condition with respect to the structure of light fields for stationary observers and local moving contrasts in the luminance function. On this basis we derive several constraints on what kind of information can be extracted from event-based sensory acquisition using the address-event-representation (AER) principle. We discuss how subsequent visual mechanisms can build upon such representations in order to integrate motion and static shape information. On this foundation we present approaches for motion detection and integration in a neurally inspired model that demonstrates the interaction of early and intermediate stages of visual processing. Results replicating experimental findings demonstrate the abilities of the initial and subsequent stages of the model in the domain of motion processing.

Keywords

neuromorphic hardware, bio-inspired modelling, event vision, motion estimation

1. INTRODUCTION

The initial stages of visual processing aim at extracting a vocabulary of relevant items related to a visual scene. This amounts to sampling the ambient optic array [Gibson, 1986] to make available behaviorally relevant structure at different hierarchical compositions. The pattern of light rays that converge to a given observer position in space and time constitute the plenoptic function $P(x, y, \lambda, t, V_x, V_y, V_z)$ [Adelson and Bergen, 1991] to define a mapping of the light-field of different intensities. One obtains a quantitative description of a volume of continuous visual measurements. As a simplification we assume a single stationary camera operating in a single narrow band of wavelengths in the electromagnetic spectrum, reducing the plenoptic function to $P_{\lambda,V_x,V_y,V_z}(x, y, t) = g(x, y, t)$ (the spatio-temporal gray level function). The pattern of light rays that constitute the optic array converge to a given observer position in space and time to communicate information about the scenic objects to the receptor surface of the viewer, or its retina. Elemental measurements are necessary to access the plenoptic structures. Conventional frame-based cameras sample the optic array reading out measurements of all light-sensitive pixels at a fixed rate. Since the temporal sampling rate is limited (sampling all pixel values in a fixed time interval) fast local luminance changes are integrated and thus lost for further processing. When no changes occur, redundant information is generated that is carried to the subsequent processing steps. Asynchronous event sensors (AER), on the other hand, employ pixels that run at individual rates generating events based on local decisions, like in the mammalian retina [Mead, 1990, Liu and Delbrück, 2010]. The primary focus in this work will be on silicon retinas that operate as AER systems (the dynamic vision sensor, DVS; [Delbruck and Liu, 2004]). Event-based sensing is frameless with local decisions made at the pixel-level based on the registration of significant changes along relevant visual dimensions. Such changes generate events $e_k \in \{-1, 1\}$ that emulate spike sequences of on- and off-contrast cells in the retina, respectively.

In this contribution, we discuss what kind of information is accessible from the initial stages of event-based visual sensing. Low- and intermediate-level visual processes build base representations in a first sweep of sensory-driven visual processing [Roelfsema, 2006]. Subsequent visual mechanisms of shape and motion processing are challenged to gain access to motion as well as to static shape information. We investigate the theoretical background underlying the initial stages of processing of an event-based spatio-temporal visual input stream. We will focus here on the detection and integration of motion information from event-based input streams and demonstrate how spatio-temporal changes in contrast can be analyzed. It has been demonstrated that fast moving

*Authors' email addresses are {stephan.tschechne; tobias.brosch; roman.sailer; nora.von-egloffstein; luma.issa; heiko.neumann}@uni-ulm.de

BICT 2014, December 01-03, Boston, United States

DOI 10.4108/icst.bict.2014.257904

shape features lead to the perceptual phenomenon of motion streaks which may augment retinal motion by so-called speed line information. We demonstrate that such motion streaks originate from fast moving structures and their integration in a temporal window defined in the form pathway. For slower speeds the integration of events deteriorates in strength and speed lines vanish. We will finally discuss the implications the event-based paradigm has for subsequent motion and form processing at intermediate and higher levels of a model cortical architecture.

2. THEORETICAL ASPECTS OF EVENT-BASED VISUAL MOTION DETECTION

To describe the output of the event based sensor, we define the function

$$e : \mathbb{R}^2 \times \mathbb{R} \rightarrow \{-1, 0, 1\} \tag{1}$$

that is always zero except for tuples $(x_k, y_k; t_k) = (\mathbf{p}_k; t_k)$ that describe the location and time of event k at which $e(\mathbf{p}_k; t_k) = e_k$ which is 1 if the log-luminance changed more than ϑ, i.e. an ON event, and -1 if it changed more than $-\vartheta$, i.e. an OFF event. This sampling of the light field leads to the temporal derivative of the luminance function g

$$\frac{d}{dt} g(\mathbf{p}; t) = g_t(\mathbf{p}; t) \approx \frac{\vartheta}{\Delta t} \sum_{t' \in (t-\Delta t, t]} e(\mathbf{p}, t') \,, \tag{2}$$

with ϑ the sensitivity threshold of the event based sensor. To estimate local translatory motion, we assume throughout the paper that the gray level function remains constant within a small neighborhood in space and time, i.e. $g(x, y; t) = g(x + \Delta x, y + \Delta y; t + \Delta t)$ (gray level constancy). Local expansion up to the second order yields the constraint $\mathbf{\Delta x}^T \nabla_3 g + 1/2 \mathbf{\Delta x}^T \mathbf{H}_3 \mathbf{\Delta x} = 0$. Here, $\mathbf{\Delta x} = (\Delta x, \Delta y, \Delta t)^T$, $\nabla_3 g = (g_x, g_y; g_t)$, and $\mathbf{H}_3$ denotes the Hessian with the 2nd order partial derivatives of the continuous gray-level function. If we further assume that the 2nd order derivative terms are negligible (linear terms dominate) we arrive at the spatio-temporal constraint equation used for least-squares motion estimation [Lucas and Kanade, 1981], i.e. $g_x u + g_y v + g_t = 0$ given that $\Delta t \rightarrow 0$ and $\mathbf{u}^T = (u, v) = (\Delta x / \Delta t, \Delta y / \Delta t)$. Note that this equation can also be solved in the frequency domain in which it holds $\omega_x u + \omega_y v + \omega_t = 0$ (with ω denoting the frequency). The local image motion $\mathbf{u}$ of an extended contrast can only be measured orthogonal to the contrast (normal flow, [Barron et al., 1994, Fermüller and Aloimonos, 1995]). For simplicity, we assume a vertically oriented gray level edge ($g_y = 0$). Then the motion can be estimated along the horizontal directions (left or right). Depending on the edge contrast polarity (light-dark, LD, $g_x < 0$ or dark-light, DL, $g_x > 0$) the spatio-temporal movements can be estimated without ambiguity. For a DL edge if $g_t < 0$ the edge moves to the right, while for $g_t > 0$ the edge moves to the left. For an LD edge the sign of the temporal derivatives g_t changes for both respective movement directions. The ratio of gray-level derivatives yields a unique direction selector, $\mathrm{sgn}(g_x / g_t) = -1$ implies rightward motion while $\mathrm{sgn}(g_x / g_t) = 1$ implies leftward motion, irrespective of the contrast polarity. A key question is to what extend the required spatio-temporal derivative information is available in the plenoptic function that is sampled by the asynchronous event sensor.

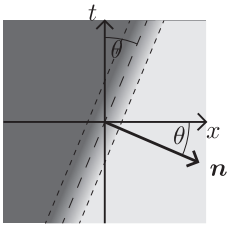

Figure 1: Rightward moving edge in 1D. The movement is described by the speed-related angle θ between the x-axis and the normal n, i.e. $\theta < 0$ in this case.

2.1 Moving gray-level edge and spatio-temporal contrast model

We describe the luminance function g for a stationary DL transition by convolving a step edge $\mathcal{H}(\cdot)$ with a parameterized Gaussian,

$$g_\sigma(x) = \frac{c}{\sqrt{2\pi}\sigma} \cdot \mathcal{H}(x) * \exp\left(-\frac{x^2}{2\sigma^2}\right) = c \cdot \mathrm{erf}_\sigma(x), \tag{3}$$

with c denoting the luminance step height and "$*$" denoting the convolution operator [Neumann and Ottenberg, 1992]. Different contrast polarities are defined by $g_\sigma^{DL}(x) = c \cdot \mathrm{erf}_\sigma(x)$ and $g_\sigma^{LD}(x) = c \cdot (1 - \mathrm{erf}_\sigma(x))$, respectively. When this transition starts moving at time $t = 0$ it generates a slanted line in the x-t-space crossing the origin (c.f. Fig. 1). The speed estimate is defined by the angle $\theta = tan^{-1}(\Delta x / \Delta t)$ that is measured against the t-axis. For a stationary gray-level edge (zero speed) we get $\theta = 0$ (i.e. the edge is located on the t-axis). Positive angles $\theta \in (0°, 90°)$ define leftward motion, while negative angles define rightward motion. Consider, for example, a DL contrast that is moving rightward (c.f. Fig. 1). The spatio-temporal gradient is maximal along the normal direction $\mathbf{n} = (\cos\theta, -\sin\theta)^T$. The function $g(x; t)$ describing the resulting picture of the movement in the x-t-space is thus given as

$$g_{\sigma\theta}(x; t) = \frac{c}{\sqrt{2\pi}\sigma} \mathcal{H}(x_\perp) * \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) \,, \tag{4}$$

with $x_\perp = x \cdot \cos\theta - t \cdot \sin\theta$. The temporal and spatial derivatives are given as

$$\frac{\partial}{\partial t} g_{\sigma\theta}(x; t) = \frac{-c}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) \cdot \sin\theta \,, \tag{5}$$

$$\frac{\partial}{\partial x} g_{\sigma\theta}(x; t) = \frac{c}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) \cdot \cos\theta \,. \tag{6}$$

Now, recall that the event-based DVS sensor provides an estimate of g_t (c.f. eqn. (2)). Thus, for a known velocity given by θ, we can combine equations (5) and (6) to determine g_x as

$$\frac{\partial}{\partial x} g_{\sigma\theta}(x; t) = -\frac{\partial}{\partial t} g_{\sigma\theta}(x; t) \cdot \tan\theta \,. \tag{7}$$

In sum, the temporal edge transition can be reconstructed from an (uniform) event sequence at the edge location for a specific motion direction. The estimation requires (i) that a reliable speed estimate is available (to get a robust value

for θ) and (ii) that temporal changes can be reliably estimated from events over an appropriately scaled temporal integration window Δw_t. The parameters θ and Δw_t need to be robustly estimated to accomplish robust estimates. The value of θ might be estimated by filtering the cloud of significant events in the space-time domain to estimate the speed and direction components of the movement. In Sec. 3 we will suggest an architecture of model cortical areas V1 and MT that utilizes filter functions and subsequent competitive interactions to extract movements of local contrast. In Sec. 2.3 we will briefly outline the necessary steps in this estimation process. Alternatively, one can directly try to estimate the partial derivatives used in the motion constraint equation from the stream of events. The construction of this approach and the related problems are described in the following section.

2.2 Estimating the spatio-temporal continuity using event-sequences

The local spatio-temporal movement of a gray-level function can be estimated by least-squares optimization from a set of local contrast measurements which define intersecting motion constraint lines in velocity space (as suggested by [Lucas and Kanade, 1981]). The key question remains how one could estimate the spatial and temporal derivatives in the constraint equations, $g_x u + g_y v + g_t = 0$ from event sequences generated by the DVS. Events only encode information about the temporal derivative g_t (c.f. Eq. (2)). Thus, without additional information it is impossible to estimate g_x or g_y. The derivative of a translatory moving gray level patch, however, generates a unique response in $h := g_t$. Thus, we can apply the motion constraint equation to the function h and solve $h_x u + h_y v + h_t = 0$, instead.

Using two temporal windows $\mathcal{T}_{-2} = (t - 2\Delta t, t - \Delta t]$ and $\mathcal{T}_{-1} = (t - \Delta t, t]$, we can approximate h_t for example by a backward temporal difference

$$h_t(\mathbf{p}; t) = g_{tt}(\mathbf{p}; t) \tag{8}$$

$$\approx \frac{\vartheta}{\Delta t^2} \left(\sum_{t' \in \mathcal{T}_{-1}} e(\mathbf{p}; t') - \sum_{t' \in \mathcal{T}_{-2}} e(\mathbf{p}; t') \right) , \tag{9}$$

with $\mathbf{p} = (x, y)^T$. The spatial derivatives h_x, h_y can be approximated by two central difference kernels $[-1, 0, 1]$ and $[-1, 0, 1]^T$, respectively. These can be applied to the function h estimated over the temporal window $\mathcal{T}$ (e.g. $\mathcal{T} = \mathcal{T}_{-2} \cup \mathcal{T}_{-1}$)

$$h_x(\mathbf{p}; t) = g_{tx}(\mathbf{p}; t) \tag{10}$$

$$\approx \sum_{t' \in \mathcal{T}} e(x + 1, y; t') - \sum_{t' \in \mathcal{T}} e(x - 1, y; t') , \tag{11}$$

$$h_y(\mathbf{p}; t) = g_{tx}(\mathbf{p}; t) \tag{12}$$

$$\approx \sum_{t' \in \mathcal{T}} e(x + 1, y; t') - \sum_{t' \in \mathcal{T}} e(x, y - 1; t') . \tag{13}$$

Obviously, the resulting flow computation results in a sparsification of responses since stationary edges will not be represented in h.

Note, however, that this approach has multiple issues in real implementations. The most severe observation is that when a bar passes the receptive field of a pixel of the DVS sensor, the amount of events is in the magnitude of about 10 events (often even smaller, depending on the contrast, speed and luminance conditions). Even for fast and strong differences the response is usually smaller than 100 events. Thus, huge approximation errors occur for h_x, h_y and especially in h_t (since this is essentially the second derivative of g). Furthermore, we can only estimate h_t accurately if the temporal windows are so small that the gray-level edge has not already passed. This limits the number of events to even less which magnifies the outlined problems.

To overcome the problem of the second derivative in time, one might think of a scheme like that proposed in [Benosman et al., 2012] in which h_x, h_y are used as estimates of g_x, g_y. As outlined in Sec. 2 this can not work in general, because without knowledge about whether it is a DL- or LD-edge, the direction of movement can not be estimated. For simple stimuli like a rotating bar or a moving belt of two types of homogeneous regions, however, this may serve as an approximation of the orientation of movement and relative speeds within the same stimulus class (c.f. [Benosman et al., 2012]).

2.3 Least-squares velocity estimation and direction-sensitive cortical filters

The short temporal window in which events of a passing contrast edge are generated makes it difficult to reliably estimate the derivatives required in the motion constraint equation (c.f. previous section). An alternative approach is to consider the distribution of events (the "event cloud") in a small volume of the x-y-t-space. The cloud resulting from a moving line generates a sandwich-like cloud defined by a plane that is oriented in x-y-t space according to the edge motion direction and speed (velocity tangent plane). The spread of the event cloud orthogonal to the velocity tangent plane (height of the "sandwich") depends on how fast the gray-level discontinuity moves through the spatial location of a pixel and its local neighborhood (the receptive field, RF, of a cell at this position). For fast motions an edge quickly passes through the RF and, consequently, only a few events are generated locally. In turn, this reduces the height of the spread with respect to the velocity tangent plane which can be inferred by a PCA or least-squares (LS) regression. In [Benosman et al., 2014] a function $\Sigma_e : \mathbb{N}^2 \to \mathbb{R}$ is defined that maps the location $\mathbf{p}$ of an event e to the time $\Sigma_e(\mathbf{p}) = t$ when it was generated. This mapping may be used to describe the cloud of events, however, it is either defined for each event in which case the mapping is not differentiable, or it is defined for all events in which case the mapping is not injective (because for a given t, there are multiple events at different locations). In any case the inverse function theorem *cannot* be applied here (disproving the approach of [Benosman et al., 2014]). Instead, the speed can be estimated from the normal vector $\mathbf{n}$ of the regression plane by solving the orthogonal system of the velocity vector $\mathbf{v}$, the orientation of the moving discontinuity edge $\mathbf{l}$, and the normal vector $\mathbf{n}$ for the speed vector. The resulting velocity components u and v are then given as (with $\mathbf{n} = (a, b, c)^T$)

$$\begin{pmatrix} u \\ v \end{pmatrix} = -\frac{c}{a^2 + b^2} \begin{pmatrix} a \\ b \end{pmatrix} , \tag{14}$$

with the speed component $s = \sqrt{u^2 + v^2} = c \cdot (a^2 + b^2)^{-1/2}$. Note, that for slow or moderate velocities, a reliable estimate of the velocity tangent plane requires a spatial as well as temporal neighborhood such that the "sandwich" height of the event cloud is fully covered within the spatio-temporal window (or RF) considered for the LS regression. If this condition is not fulfilled, i.e. if the window is smaller than the extend of the cloud, then the principal axes are arbitrary and cannot be estimated reliably. As an alternative to considering the LS regression in estimating the velocity tangent plane from the cloud of events the uncertainty of the event detection might be incorporated directly. At each location detected events define likelihood distributions given certain velocities, $p(e|\mathbf{u})$. A maximum likelihood estimate might be defined to detect the velocity of a visual structure in the light field using Bayesian inference. Based on current knowledge of the filter characteristics of V1 spatio-temporally cells, we describe a related scheme using filter mechanisms in Sec. 3.1. A competitive interaction between filter responses realizes a maximization mechanism in neural architecture.

3. EVENT-BASED MOTION ESTIMATION

In the following we build upon a biologically inspired approach for motion estimation in the early stages of visual cortex using the input of an event-based vision sensor [Tschechne et al., 2014]. Our contribution contrasts to approaches like [Benosman et al., 2012] that adapted a gradient-based [Lucas and Kanade, 1981] approach using error minimization. We on the other hand model a simplified version of the initial stages of cortical visual processing along the dorsal and ventral pathway in primates. Our proposal works on the continuous stream of events using motion energy selective filters in a feed-forward sweep of hierarchical processing, utilizing increasingly larger receptive fields to build feature representations of higher complexity [Ungerleider and Haxby, 1994]. In particular, initial responses are generated to represent movements in the spatio-temporal domain, corresponding to V1 direction sensitive cells. These activations are integrated at model MT and provide modulatory feedback for V1 cells and by this improve the representation of motion. This employs a three-stage cascade of initial processing, feedback entry and normalization, that builds upon earlier studies of motion and form processing, e.g.[Neumann and Sepp, 1999].

3.1 Filter design and spatio-temporal motion detection

When motions occur in the visual field of the sensor, the sampled events are generated at the contrast edges of the moving object. In x-y-t-space these events cluster into a region where events occur with a high probability and this region is slanted with respect to the t axis, where the angle indicates the speed of the motion, see Sec. 2.3 and Fig. 3. Filters selective to motion thus need to be selective to such a slanted region of events. The design of the spatio-temporally inseparable filters used in our model incorporates evidence that they can be generated by superposition of temporal and spatial components [De Valois et al., 1982]. Cells in V1 split into populations showing responsiveness to direction or not. While the non-directionally selective cells also contribute to the motion estimation (see Sec. 4), we first describe the design of the directionally-selective cells. Those are separable with respect to time and space and thus can be generated by means of a combination of simpler components. While spatial components resemble the characteristic of oriented even- and odd-symmetric Gabor functions, the temporal components distribute into those with a biphasic temporal component and those with a monophasic temporal component. These components interact in a specific scheme to yield a spatio-temporally tuned filter that will react on patterns that move in the x-y-t-domain. This slightly contrasts with [Adelson and Bergen, 1985] who used a different set of generating functions. Two of these components (temporally bi-phasic, spatially even and temporally monophasic, spatially odd, respectively) are combined at a time to already build spatio-temporal selective cells. Parameters of the filters contribute to the speed- and direction selectivity of the resulting filter. In the following, we describe the filter functions used in our model. Events are buffered into a first-in-first-out data structure to allow quick lookup of neighboring events to estimate their relative integration weight. Spatial filters are modeled using a population of rotated two-dimensional Gabor functions with isotropic Gaussian window:

$$G(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\pi\left[\frac{(\hat{x}-x_0)^2}{\sigma^2} + \frac{(\hat{y}-y_0)^2}{\sigma^2}\right]\right) \cdot exp\left(i[\xi_0 x + \xi_0 y]\right) \tag{15}$$

where $\begin{pmatrix}\hat{x} & \hat{y}\end{pmatrix}^T = R_\theta \cdot \begin{pmatrix}x & y\end{pmatrix}^T$ with R_θ defining the planar rotation. Even and odd components are represented in real ($\mathbb{R}$) and imaginary ($\mathbb{I}$) components of this function and parametrized to generate 11×11 filter kernels ($x, y \in [-\pi..\pi]$) using standard deviation $\sigma = 2.5$, frequency tuning $\xi_0 = 2\pi$ and rotation $\theta = [0, \frac{1}{4}\pi, .., \pi]$. Fig. 3 depicts some of the filters used. Temporal filter functions are generated using the following equation:

$$f(t)_{mono/bi} = w_{m1} \cdot exp\left(-\frac{(t-\mu_{m1})^2}{2\sigma_{m1}^2}\right) - w_{m2} \cdot \exp\left(-\frac{(t-\mu_{m2})^2}{2\sigma_{m2}^2}\right) \tag{16}$$

For monophasic temporal kernels f_{mono} these values are set to $w_{m1} = 1.95$, $\mu_{m1} = \mu_{m2} = 0.55$, $\sigma_{m1} = 0.10$, $w_{m2} = 0.23$, $\sigma_{m2} = 0.16$. For biphasic temporal kernels f_{bi} the values are set to $w_{m1} = 0.83$, $\mu_{m1} = 0.44$, $\sigma_{m1} = 0.12$, $w_{m2} = -0.34$, $\mu_{m1} = 0.63$, $\sigma_{m2} = 0.21$. Due to the sparse representation of filter responses, a full convolution of the input events with the filter kernels is not necessary. We incorporate a normalization stage at the output of the feed-forward filtering and top-down feedback modulation cascade:

$$\frac{\partial}{\partial t} r^{V1}_{\theta\mathbf{p}} = -\alpha \cdot r^{V1}_{\theta\mathbf{p}} + I_{ex} \cdot (1 + \beta \cdot r^{MT}_{\theta\mathbf{p}}) - \gamma \cdot r^{V1}_{\theta\mathbf{p}} \cdot q^{pool}_{\theta\mathbf{p}}, \tag{17}$$

$$\frac{\partial}{\partial t} q^{pool}_{\theta\mathbf{p}} = -q^{pool} + \sum_{\phi,\mathbf{p}'} r^{V1}_{\phi\mathbf{p}'} \cdot \Lambda^{pool}_{\theta\phi\mathbf{p}\mathbf{p}'}, \tag{18}$$

with γ denoting the strength of divisive inhibition. The multiplicative term $1 + \beta \cdot r^{MT}$ denotes the response gain modulation generated by top-down feedback generated by model MT cells [Bayerl and Neumann, 2004]. In case of a pure feed-forward signal processing sweep, the modulation is switched off by setting $\beta = 0$. The bottom-up input filter

response is generated by

$$I(x,y,t)_{ex} = \sum_{i,j,t'} e(x,y,t) \cdot [f(t)_{bi} \cdot \mathbb{R}(G(x,y)) + f(t)_{mono} \cdot \mathbb{I}(G(x,y))] \quad (19)$$

In a purely feed-forward sweep $\beta = 0$ (compare Sec.3.3).

3.2 Results of initial motion estimation

In the following we demonstrate how the approach robustly estimates optic flow for a set of test stimuli. The tests utilize simple stimuli of translatory and rotational motion with known ground truth. Results are shown in Fig. 2. We integrated the results of a time span to be visualized in this paper, because single motion events cannot sensibly be presented in printed form. We chose an integration window of 50ms. Where adequate we calculated an error measure for our results as follows. With the type of motion known, we synthesize a ground truth vector field of a linear or rotational motion for each sequence. The estimated error is the angular error between zero and 180 degrees between the synthetic vector field from ground truth and the estimated motion direction at this position.

3.3 Motion integration and re-entry

In the previous section we demonstrated how direction-selective neurons in model area V1 may encode spatio-temporal changes of visual patterns. Such cells respond coarsely to movements of gray-level structures given a direction ϕ orthogonal to their orientation selectivity θ and over a broad range of speeds. Such responses, denoted by $r^{V1}_{\theta,\phi}(x,y,t)$, are integrated by cells in area MT which obey an increased selectivity to direction and speed [Born and Bradley, 2005]. We incorporated a stage of integrating early motion responses from model V1 using circular receptive field weighting functions Λ with Gaussian profile but larger spatial integration size over a neighborhood approximately 5 times the size of V1 filters. The responses are formally calculated by the following mechanism (see [Brosch and Neumann, 2014] for details of the generic circuit model):

$$\partial_t r^{MT}_{\phi,s} = -\alpha r^{MT}_{\phi,s} + \sum_{i,j,\theta} r^{V1}_{\theta\phi}(i,j) \cdot \Lambda_{xy,ij} \Phi_s - r^{MT}_{\phi,s} \cdot q^{MT} \quad (20)$$

(for better readability we omitted space-time locations as parameters where possible). The activations r represent membrane potentials and are assumed here to correspond to firing rates in a one-to-one fashion (thus, assuming a linear transfer function). In a nutshell, the mechanism integrates V1 responses with a specific spatio-temporal selectivity over a larger spatial and temporal neighborhood to generate a velocity selectivity as specified by Φ_s. We also apply a down-sampling operator to the resulting representation of $r^{MT}_{\phi,s}(x,y,t)$. A pool normalization is incorporated that acts divisively upon the activities. The mechanism is formally defined as

$$\partial_t q^{MT} = -q^{MT} + \beta \sum_{i,j,(\phi,s)} r^{MT}_{\phi,s}(i,j) \cdot \Lambda^{pool}_{xy,ij}. \quad (21)$$

This stage sums the responses over a neighborhood in the spatio-velocity domain. These responses are used as feedback r^{MT} as introduced in Eq. 17.

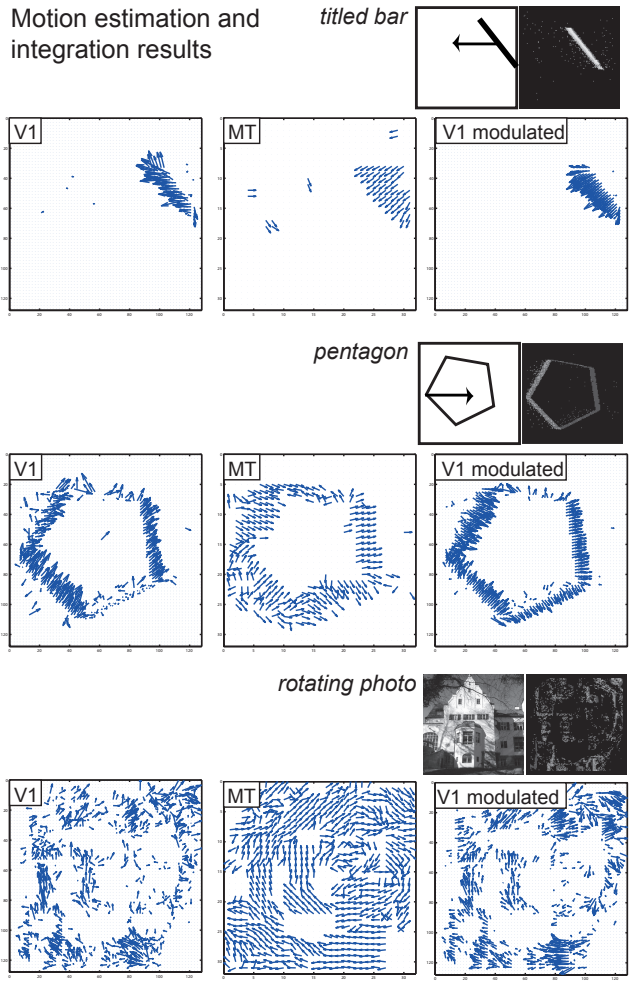

Figure 2: Results of motion estimation and integration. Stimulus, motion and event representation in small images. Initial motion estimation, representation at model MT, V1 representation with applied feedback in large images (from left to right).

Fig. 2 shows examples of the effect the modulatory feedback has. The feedback improves the estimation of normal flow (first two examples) while it simultaneously smooths the representation of motion at V1 level (third example.)

4. MOTION STREAKS

In the previous section we demonstrated how direction-selective neurons in model area V1 may encode spatio-temporal changes of visual patterns. Such cells respond coarsely to movements of gray-level structures making use of non-separable filters that are oriented in the space-time domain. Can non-directionally tuned cells also contribute to the estimation of motion direction? Evidence from perceptual psychophysical investigations suggest that the visual system utilizes "motion streak" responses as a spatial code for local motion direction [Geisler, 1999, Burr, 2000]. Such encoding might preferentially use the form channel [Ungerleider and Haxby, 1994] where they form oriented streaks, or speed lines [Apthorp et al., 2013]. Those streaks become apparent only for fast movements of structured scenic patterns, while they are suppressed under normal vision conditions [Wallis and Arnold, 2009]. They are oriented parallel to the available motion direction and would aid determining the visual motion direction. The neural mechanisms underlying the detection and integration of such motion streaks in the form channel representation remain largely unclear. We argue that the formation of responses that relate to motion streaks can be naturally explained by the spatial integration of, e.g., grouping cells in the form pathway, that integrate their input over an approximately fixed temporal window. For slow to moderate pattern speeds only a few input events will be integrated in such grouping cells within the given temporal window. However, for fast speeds appropriately oriented integrations cells, as in area V2 [Peterhans and von der Heydt, 1991] (see [Neumann et al., 2007] for model definitions), will integrate several incoming events within the given time window. As a result, such cells will elicit a response in an orientation that corresponds to the motion direction. In the following we propose how a representation of motion streaks and an interaction with processes of motion estimation could be explained by means of interactions of early visual areas. The principles are modeled using the event-based representation. Sec. 3 presented a model using spatio-temporal filters at V1 to detect visual motion, that is integrated in MT. We extend this model by incorporating model V2 cells that are spatially tuned and respond to oriented structures. Such cells integrate recent visual events over a fixed temporal window. When sufficiently fast motion is presented, these cells integrate enough events along their preferred orientation to become activated. While this only occurs for fast motions along the preferred orientation, the activation would not be generated for slow motions. We employ a simplified variant of such orientation selective cells. Instead of bi-lobed figure-eight-shaped grouping filters (as discussed in [Neumann et al., 2007]) we use elongated weighting functions with Gaussian weights $K^{group,V2}_{\theta,\tau}(x,y,t)$. The elongation of the weighting is defined by the ratio between long and short axis, $\tau = \sigma_1/\sigma_2$. The orientation selectivity is defined by the proper rotation, R_θ. The responses of model cells are defined in steady state by

$$r^{V2}_\theta = \sum_{i,j} r^{V1}_\theta \cdot K^{group,V2}_{\theta,\tau}(x-i, y-j, t) \qquad (22)$$

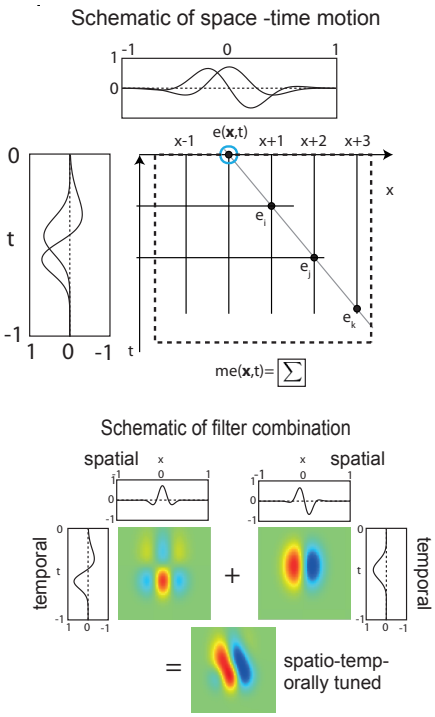

Figure 3: ***Top:*** **In an event based representation, an object leaves a trail of events when moving across the visual array, with the structure slanted to the t axis as a function of speed.** ***Middle:*** **Spatio-temporal filters are generated using two spatially and two temporally tuned functions, motivated by physiological findings.**

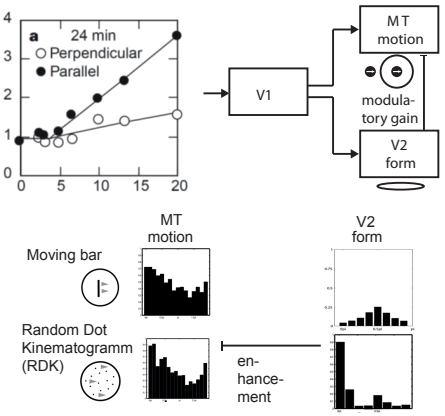

Figure 4: ***Top left:*** **Fast motions leave a trail of activation in the spatio-temporal domain. Their detection in the form channel contributes to motion estimation by providing a spatial code for motion direction.** ***Top right:*** **Model structure.** ***Bottom:*** **Estimation and enhancement results for two stimuli.**

Our model contains cells tuned to orientation using an elongated Gaussian kernel (including a planar rotation):

$$G_\theta(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\pi\left[\frac{(\hat{x}-x_0)^2}{\sigma_1^2} + \frac{(\hat{y}-y_0)^2}{\sigma_2^2}\right]\right) \quad (23)$$

The interaction between motion direction cells tuned to estimated motion in direction ϕ and cells oriented along that direction is modelled using a modulatory interaction. We employ a response modulation mechanism as investigated in [Brosch and Neumann, 2014], such that

$$r_{\phi,s}^{MT} \propto r_{\theta,\phi}^{V1} \cdot (1 + r_\theta^{V2}) \quad (24)$$

This interaction requires a pool normalization as already introduced in Eq. 21. We tested our model with a stimulus consisting of random dot patterns to replicate experimental settings, see Fig. 4. Results indicate that likewise oriented contrast-sensitive cells are co-activated parallel to motion direction, thus representing motion streaks. Without the need to assume separate motion channel representations, these streaks occur in the form channel as a direct consequence of fast coherent motions along single directions. This suggests an interaction of motion an form, while stipulating the prediction that the coherence of motion direction and streak orientation is only available for fast motions of random dot kinematrograms, and less available for moving edges. Here, the oriented integration field would show highest activation for orthogonally tuned orientation.

5. SUMMARY AND DISCUSSION

This manuscript investigates motion estimation in the light of event-based input generation and representation. This method of sampling the plenoptic function is fundamentally different from common frame-based approaches of image acquisition. Here, local samples are generated by individual sensory elements (pixels) which are decoupled from a global synchronization signal. So far, a small number of papers have investigated how classical approaches in computer vision can be adapted to event-based sensory input and how the quality of the results changes depending on the new data representation framework. Examples are [Benosman et al., 2012, Benosman et al., 2014] for optical flow computation and [Rogister et al., 2011, Piatkowska et al., 2013, Camunas-Mesa et al., 2014] for stereo vision. Furthermore, other authors show future applications of this new sensor technology that has the potential to provide fast, robust and highly efficient sensory processing in various domain and challenging scenarios [Fu et al., 2008, Lichtsteiner et al., 2008, Drazen et al., 2011]. Our focus is on motion computation and the proposed approach is different from previous approaches in several respects. We first investigate fundamental aspects of the local structure of light fields for stationary observers and local moving contrasts in the luminance function. On this basis we derive several constraints on what kind of information can be extracted from event-based sensory acquisition using the AER principle. Based on these insights several previous approaches can be unified into a common framework of event-based motion detection. We then propose a hierarchical architecture of multi-stage motion detection and integration that builds upon experimental findings of the primate visual system. In particular, we propose a filter approach that samples spatio-temporal luminance changes and generates a representation of oriented moving contrasts, like in cortical area V1. Because such initial estimates are still rather noisy and provide only a rough estimate of the stimulus velocity, these initial estimates are integrated by mechanisms that relate to cortical area MT. They sum multiple spatio-temporal responses over a larger neighborhood, irrespective of their orientation, but sensitive to velocities. This is, to our knowledge, the first event-based scheme with realistic multi-stage estimation of motion that is based on sparse event-based input. As such it is inspired by previous approaches of (frame-based) motion detection [Bayerl and Neumann, 2004, Raudies et al., 2011] and extends a proposal of initial motion detection suggested in [Tschechne et al., 2014]. In addition, we close the loop incorporating re-entrant signals from model area MT back to V1 using a previously developed modulatory feedback scheme. Such feedback is part of a columnar model of recurrent inter-areal processing such that higher level representations can enhance and, thus stabilize, representations derived from sparse and noisy input. Details of this model have been developed before [Bouecke et al., 2011, Brosch and Neumann, 2014]. The model proposed here shows that the primary role of the feedback at such early signal related stages might indeed be the enhancement of coherent input signals and reduction of noise and spurious responses [Pratte et al., 2013]. In addition to this recurrent scheme of motion detection and integration we also propose how an integration of inputs by grouping cells in V2 may generate responses that lead to spurious representations of so-called motion streaks, or speed-line representations [Geisler, 1999, Apthorp et al., 2013]. We show how this mechanism automatically follows from high temporal resolution of input event generation and the integration of such events along certain extended grouping orientations in the form system. We further propose a simple scheme of modulating interaction between form representations and motion responses in model MT. We demonstrate how motion estimation can be improved by such interactions from the form channel where form representations along motion traces can improve and stabilize their representation.

6. ACKNOWLEDGEMENTS

This work has been supported by grants from the Transregional Collaborative Research Center SFB/TRR62 funded by the German Research Foundation (DFG). LIAK has been supported by grants from the Ministry of Higher Education and Scientific Research (MoHESR) Iraq and from the German Academic Exchange Service (DAAD).

7. REFERENCES

[Adelson and Bergen, 1991] Adelson, E. H. and Bergen, J. H. (1991). The plenoptic function and the elements of early vision. In Landy, M. and Movshon, J. A., editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press.

[Adelson and Bergen, 1985] Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.*, 2(2):284–299.

[Apthorp et al., 2013] Apthorp, D., Schwarzkopf, D. S., Kaul, C., Bahrami, B., Alais, D., and Rees, G. (2013). Direct Evidence for encoding of motion streaks in human visual cortex. *Proc R Soc B*, 400(20122339).

[Barron et al., 1994] Barron, J., Beauchemin, S., and DJ, F. (1994). Performance of optical flow techniques. *Int'l*

J. of Computer Vision, 12(1):43–77.

[Bayerl and Neumann, 2004] Bayerl, P. and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16(10):2041–66.

[Benosman et al., 2014] Benosman, R., Clercq, C., Lagorce, X., Ieng, S. H., and Bartolozzi, C. (2014). Event-Based visual flow. *IEEE Trans. on Neural Networks and Learning Systems*, 25(2):407–417.

[Benosman et al., 2012] Benosman, R., Ieng, S.-H., Clercq, C., Bartolozzi, C., and Srinivasan, M. (2012). Asynchronous frameless event–based optical flow. *Neural Networks*, 27:32–37.

[Born and Bradley, 2005] Born, R. and Bradley, D. (2005). Structure and function of visual area MT. *Annu. Rev. Neurosci.*, 28:157–189.

[Bouecke et al., 2011] Bouecke, J. D., Tlapale, E., Kornprobst, P., and Neumann, H. (2011). Neural mechanisms of motion detection, integration, and segregation: from biology to artificial image processing systems. *Eurasip Journal on Advances in Signal Processing*.

[Brosch and Neumann, 2014] Brosch, T. and Neumann, H. (2014). Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Networks*, 54:11–6.

[Burr, 2000] Burr, D. C. (2000). Motion vision: are 'speed lines' used in human visual motion? *Curr Biol.*, 10:R440–443.

[Camunas-Mesa et al., 2014] Camunas-Mesa, L. A., Serrano-Gotarredona, T., Ieng, S. H., Benosman, R., and Linares-Barronco, B. (2014). On the use of orientation filters for 3D reconstruction in event-driven stereo vision. *Frontiers in Neuroscience*, 8(48).

[De Valois et al., 1982] De Valois, R. L., Albrecht, D. G., and Thorell, L. G. (1982). Spatial Frequency Selectivity of cells in macaque visual cortex. *Vision Research*, 22:545–559.

[Delbruck and Liu, 2004] Delbruck, T. and Liu, S. (2004). A silicon early visual system as a model animal. *Vision Research*, 44:2083–2089.

[Drazen et al., 2011] Drazen, D., Lichtsteiner, P., Häflinger, P., Delbrück, T., and Jensen, A. (2011). Toward real-time particle tracking using an event-based dynamic vision sensor. *Exp. Fluids*, 51:1465–1469.

[Fermüller and Aloimonos, 1995] Fermüller, C. and Aloimonos, Y. (1995). Qualitative egomotion. *Int'l J. of Computer Vision*, 15(1–2):7–29.

[Fu et al., 2008] Fu, Z., Delbrück, T., Lichtsteiner, P., and Culurciello, E. (2008). An address-event fall detector for assisted living applications. *IEEE Trans. on Biomedical Circuits and Systems*, 2(2):88–96.

[Geisler, 1999] Geisler, W. S. (1999). Motion streaks provide a spatial code for motion direction. *Nature*, 400:65–69.

[Gibson, 1986] Gibson, J. (1986). The Ecological Approach to Visual Perception. *Hillsdale, NJ, Lawrence Erlbaum Associates*.

[Lichtsteiner et al., 2008] Lichtsteiner, P., Posch, C., and Delbrück, T. (2008). A 128×128 120dB 15μs Latency Asychronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid–State Circuits*, 43(2):566–576.

[Liu and Delbrück, 2010] Liu, S. C. and Delbrück, T. (2010). Neuromorphic Sensory Systems. *Current Opinion in Neurobiology*, 20:288–295.

[Lucas and Kanade, 1981] Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. IJCAI, pages 674–679.

[Mead, 1990] Mead, C. (1990). Neuromorphic electronic systems. *Proc. IEEE*, 78(10):1629–1636.

[Neumann and Ottenberg, 1992] Neumann, H. and Ottenberg, K. (1992). *EUSIPCO-92: Theories and Applications*, volume I, chapter Estimating ramp-edge attributes from scale-space, pages 603–607. Elsevier.

[Neumann and Sepp, 1999] Neumann, H. and Sepp, W. (1999). Recurrent v1-v2 interaction in early visual boundary processing. *Biological Cybernetics*, 81:425–44.

[Neumann et al., 2007] Neumann, H., Yazdanbakhsh, A., and Mingolla, E. (2007). Seeing surfaces: The brain's vision of the world. *Physics of Life Rev.*, 4:189–222.

[Peterhans and von der Heydt, 1991] Peterhans, E. and von der Heydt, R. (1991). Subjective contours – bridging the gap between psychophysics and physiology. *Trends in Neuroscience*, 14(3):112–119.

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

[Pratte et al., 2013] Pratte, M. S., Ling, S., Swisher, J. D., and Tong, F. (2013). How attention extracts objects from noise. *J Neurophysiol*, 110:1346–1356.

[Raudies et al., 2011] Raudies, F., Mingolla, E., and Neumann, H. (2011). A model of motion transparency processing with local center-surround interactions and feedback. *Neural Computation*, 23(11):2868–914.

[Roelfsema, 2006] Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annual Review of Neuroscience*, 29:203–227.

[Rogister et al., 2011] Rogister, P., Benosman, R., Ieng, S. H., and Posch, C. (2011). Asynchronous event–based binocular stereo matching. *IEEE Transactions on Neural Networks*, 22(11):1723–1734.

[Tschechne et al., 2014] Tschechne, S., Sailer, R., and Neumann, H. (2014). Bio-inspired optic flow from event-based neuromorphic sensor input. In *ANNPR 2014*.

[Ungerleider and Haxby, 1994] Ungerleider, L. G. and Haxby, J. V. (1994). 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4(2):157–65.

[Wallis and Arnold, 2009] Wallis, T. and Arnold, D. (2009). Motion-induced blindness and motion streak suppression. *Curr Biol.*, 19(4):325–329.