ulm university universität
uulm

Neural Mechanisms of Visual Perception Using Address-Event Representation

DISSERTATION

zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS

(DR.-ING.)

der Fakultät für Ingenieurwissenschaften,
Informatik und Psychologie der Universität Ulm

von

Luma Issa Abdul-Kreem Al-Helli

aus Baghdad/Iraq

Betreuer: Prof. Dr. Heiko Neumann
Prof. Dr. Dr.-Ing. Wolfgang Minker

Amtierende Dekanin: Prof. Dr. Tina Seufert

Ulm, 21.3.2016

Amtierende Dekanin der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie:
Prof. Dr. Tina Seufert

Gutachter: Prof. Dr. Heiko Neumann
Prof. Dr. Dr.-lng. Wolfgang Minker

Prüfer: Prof. Dr.-lng. Albrecht Rothermel
Prof. Dr. Karl-Joachim Ebeling

Datum der Einreichung: 21. 3. 2016

Abstract

The visual cortex analyzes motion information along a hierarchical architecture composed of several visual areas. These areas interact through bidirectional interconnections in which bottom-up connections pass on the driving visual input from lower areas to higher areas, while top-down connections re-enter higher level representations to the earlier areas via feedback. This thesis proposes a bio-inspired visual model for motion detection and integration in which feedforward and feedback interactions are incorporated. Our model takes into account two main areas in the visual cortex along motion pathway, V1 and MT, with different feature selectivities, V1 (direction selectivity) and MT (velocity selectivity). The model highlights how the visual cortical areas interact and communicate with each other based on the different representations involved. A neuromorphic vision sensor (silicon retina) is exploited where the input is generated by an event sensor that is sensitive to temporal changes in the input luminance. The neuromorphic vision sensor simulates the spike-generation functionality of the biological retina which is unlike a frame-based sensor that generates fully integrated luminance functions. This thesis shows how the functionality of Adelson-Bergen spatiotemporal filtering is adapted to be consistent with the bio-mimetic architecture of the silicon retina. The model uses a normalization method that operates upon a pool of cells in the surrounding neighborhood of spatial and feature selectivity. This normalization reduces the overall activation and tends to conserve the overall signal energy in the pool of cells which interact competitively. The prior enhancement of selected cells in turn reduces the activity of those cells that have not received any feedback. The model makes use of modulatory feedback mechanisms, which enhance driving feedforward activations. In order to show the robustness of our model, we examined the model by probing it with synthetically generated ground truth stimuli and realistic complex motions. The results reveal that the error margin between the estimated motion and synthetic ground truth is decreased in area MT compared with the initial estimation of area V1. The modulated V1 cell activation shows an enhancement of the initial motion estimation that is steered by feedback signals from MT cells. In addition, we

demonstrate the usability of the suggested model for classification purposes where action recognition is selected as a case study. The results highlight the richness of the features extracted from our model which leads to an accurate recognition of biological movements acquired from human articulations.

Acknowledgment

I would like to thank my supervisor Prof. Heiko Neumann for his unlimited support and help. I benefited from his advice, comments and his careful editing that contributed enormously to the production of paper manuscripts. I would like also to thank Prof. Günther Palm, who kindly accepted to write several referee-letters for DAAD and for his support during the course of my study.

I am very grateful to the following institutions for offering me grants during my research: Ministry of Higher Education and Scientific Research (MoHESR) Iraq, the German Academic Exchange Service (DAAD) and the international office at Ulm University (DAAD).

I thank all our vision group members specially my colleague Stephan Tschechne for all his help. Also, I would like to thank the institute of Neural Information Processing for good hosting. I believe that teamwork is the key to success in Ph.D. work, "It takes two to tango".

Finally, my thanks go to all my family members, who have never stopped encouraging me. Most importantly, my deepest thanks go to my husband.

Contents

Abstract i

Acknowledgment iii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Visual perception . . . 2
  - 1.2.1 General physiology . . . 3
  - 1.2.2 Overview of the visual areas related to this work . . . 5
    - 1.2.2.1 Neuron properties in areas V1 and MT . . . 8
    - 1.2.2.2 Motion perception and aperture problem . . . 9
- 1.3 Neural modeling . . . 11
- 1.4 Motion detection . . . 14
  - 1.4.1 Gradient techniques . . . 15
  - 1.4.2 Frequency-based techniques . . . 16
- 1.5 Outline of the thesis . . . 20

2 Motion Detection 23
- 2.1 Introduction . . . 23
- 2.2 DVS . . . 24
  - 2.2.1 Address-Event Representation (AER) . . . 25
  - 2.2.2 Input representation of ON/ OFF events . . . 28
- 2.3 Experimental devices . . . 30
- 2.4 Event-based . . . 30
  - 2.4.1 Motion estimation based on AER principles . . . 33
    - 2.4.1.1 Gradient based-event representation . . . 33
    - 2.4.1.2 Filtering based-event representation . . . 34
- 2.5 Bio-inspired model . . . 35
  - 2.5.1 Initial input stage (silicon retina) . . . 37
  - 2.5.2 Motion detection stage (area V1) . . . 38
    - 2.5.2.1 Spatial filters . . . 38

2.5.2.2 Temporal filters . . . 39
2.5.2.3 Combination of spatial and temporal filters . . 40
2.5.3 Response normalization . . . 42
2.6 Experimental setup and results . . . 45
2.6.1 Ground truth data . . . 45
2.6.2 Complex realistic movements . . . 48
2.7 Aperture problem . . . 49
2.7.1 Intersection of Constraints and Vector Average . . . 51
2.7.2 Motion integration based on biologically plausible . . . 52
2.8 Discussion . . . 52

3 Neural Mechanisms 57
3.1 Introduction . . . 57
3.2 Motion processing architecture . . . 58
3.2.1 Motion estimation-feedforward processing . . . 59
3.2.1.1 Model mechanism for area MT . . . 59
3.3 Response normalization . . . 64
3.4 Feedback from area MT to V1 . . . 68
3.5 Results . . . 69
3.5.1 Speed selectivity in areas V1 and MT . . . 70
3.5.2 Motion representation in feedforward and feedback sweeps 71
3.5.2.1 Feedforward motion representation . . . 71
3.5.2.2 Feedback motion representation . . . 84
3.6 Discussion . . . 86
3.6.1 Summary and main contributions . . . 86
3.6.2 Feedforward and feedback interaction . . . 89
3.6.3 Relation to previous models of motion estimation . . . 90
3.6.4 Model evaluation and future work . . . 94

4 Articulated motion 95
4.1 Introduction . . . 95
4.2 Human action recognition . . . 96
4.2.1 Previous work . . . 96
4.3 Our model . . . 98
4.3.1 Feature vector . . . 98
4.3.2 Experimental protocol . . . 101
4.3.3 Supervised learning of classifier . . . 101
4.3.4 Experimental setup and results . . . 103
4.3.5 Summary and Discussion . . . 107

5 Summary **109**
5.1 Main contributions 110
5.2 Relevant publications 113

List of Figures

1.1 The hierarchy of the visual areas 4
1.2 Receptive field structure of the neurons in areas LGN and V1 . 7
1.3 Receptive field structure of the MT neurons 8
1.4 Spatiotemporal frequency plots 10
1.5 Examples for five levels of single-cell modeling 13
1.6 Motion representation of a black box 17
1.7 Motion energy model 19

2.1 The Dynamic Vision Sensor silicon retina. 26
2.2 Examples of the DVS camera output for two stimuli. 27
2.3 The output of the DVS128 camera. 29
2.4 The experimental devices that used in our recording process. . . 32
2.5 Overview of the proposed model structure. 36
2.6 Initial input stage. 37
2.7 Fast and slow temporal filters. 40
2.8 Spatiotemporal filter construction. 41
2.9 The stages of our suggested model following to the Adelson-Bergen energy model. 43
2.10 Processing results of translatory motion stimuli, straight-bar and star. 46
2.11 Processing results of clockwise rotational motion stimuli 47
2.12 Estimated motion of bouncing ball. 50
2.13 Estimated flow of articulated motion of an actor. 51
2.14 Aperture problem 53

3.1 Feedforward and feedback neural model structure. 60
3.2 Block diagram of V1-MT feedforward and feedback processing. . 61
3.3 speed representation in spatiotemporal (x,t) domain 63
3.4 The model of area MT. 65
3.5 Bar movement with three different speeds. 72
3.6 The distribution of V1 responses. 73

3.7 MT cell selectivity in the spatiotemporal (x, t) domain. 74
3.8 Motion estimation of tiger and ball stimuli. 76
3.9 Motion estimation of bars movement through circular aperture. 78
3.10 Motion estimation of bars movement through rectangular aperture 79
3.11 Motion estimation of 2D structure of corner terminators 81
3.12 Motion estimation for smooth-cross and plaid stimuli. 82
3.13 Speed selectivity of MT cells. 83
3.14 Response distribution of the speed selective MT cells. 85
3.15 The distribution of area V1 responses and the modulated V1 responses. 87

4.1 Focus on human movement. 99
4.2 Motion dynamics detection of human activity. 100
4.3 Biological motions. 102
4.4 Confusion matrix based on feature extraction of 0.10s 104
4.5 Confusion matrix based on feature extraction of 0.15s. 104
4.6 Confusion matrix based on feature extraction of 0.20s. 105
4.7 The error recognition rate of the test set 106

Chapter 1

Introduction

1.1 Motivation

Visual perception enables interaction with the surroundings and is thus key ability in the field of biology, and human beings in particular. The human visual system captures visual information from the outside world through sensory transduction which encodes this information to neural signals and sends them across pathways that interpret the visual information. The analysis of visual motion is a crucial aspect of our existence which permeates all of our activities enabling us to predict and understand moving objects . Humans can easily understand visual scenes that consist of translatory, rotatory, articulated or biological motions such as walking, running, jumping, and so on. These motions require sophisticated computations in order to be interpreted and analyzed by our brain. The rapid and high accuracy of our visual system highlights the superiority of our brain compared with even the most complicated artificial cognitive systems.

Over the last few decades, investigations in the fields of cognitive psychology and neuroscience have increasingly led to a greater understanding of the underlying mechanisms of the brain. These studies have revealed much of how the brain functions and of how it is influenced and activated via a stimulus, thereby revealing that the brain is a powerful neurocognitive system. Many details of the brain's visual areas V1 and MT have been investigated, for example, how neurons in areas V1 and MT respond to stimulus and what the properties are of the neurons in both areas. However, several issues still need to be examined, such as how the visual areas can interact and how high level areas can project feedback signals to low level areas where each has particular properties.

The evolution of studies in visual processing has involved several disci-

plines, for example, neurosciences, computer sciences, mathematics, image processing, and electrical engineering. This interdisciplinary approach has increase the number of simulation models of the brain's cortical areas that can be used to guide automatic behavior for many applications such as machine vision, surveillance, object recognition and biological motion detection. Developing hardware and software in parallel has paved the way for artificial systems with high performance, more accuracy, and reduced computational time. In artificial vision systems, vision sensors are often used to equip visual systems with visual information. One of the promising techniques to emulate biological retina sensing is the neuromorphic vision sensor. Such systems use address-event representation (AER), in which each pixel of the vision sensor responds instantaneously and independently based on intensity changes. The sensor generates spikes of events with sub-millisecond timing precision, through which subtle and rapid motion can be detected. Neuromorphic vision sensors are increasingly being used in different applications, however, few models for motion estimation have been reported on how to build an adaptive model that is consistent with the functionality of neuromorphic vision sensors.

Motion estimation research is a burgeoning field and has various interesting aspects. Our aim was to build a bio-inspired model for motion estimation that simulates the computational process of the brain's areas. This model has a realistic multi-stage estimation of motion which considers the cortical areas that are dedicated for motion estimation. The model uses neuromorphic vision sensors for data acquisition rather than a conventional frame-based camera. This sensor equips the model for data acquisition with a high temporal resolution and reduces the redundant information captured by a frame-based camera. Such a visual sensory system requires an adaptive model which is consistent with the functionality of AER principles. The model can be used in various applications such as object recognition and biological motion detection.

1.2 Visual perception

In visual perception, motion is a significant source of information for the cortical visual system where this information is analyzed and processed via a considerable number of interconnected areas [Felleman and Van Essen, 1991]. The visual input is evaluated and processed by a hierarchy of brain areas where each area contains layers of neurons that are characterized by specific properties [Lamme et al., 1998]. The neuron responses increase in complexity from the lower level to the higher cortical area [Perrone, 2004]; [Ungerleider and Haxby, 1994].

Along the visual cortex hierarchy, the visual input is evaluated and inter-

preted into contextual information through two streams, namely the dorsal and ventral pathways [Mishkin et al., 1983]; [Ungerleider and Mishkin, 1982]. The dorsal pathway (also known as motion or where stream) is involved in the analysis of visual motion information where this pathway propagates from V1 in the occipital lobe toward the parietal lobe (areas MT, MST). In parallel with dorsal pathway, a progression of areas from V1, V2, and V4 running ventrally toward the temporal lobes (also known as form or what stream) is specialized for the analysis of visual attributes and object recognition other than motion [Bear et al., 2007]. Figure 1.1 shows the hierarchy of the visual areas in which 1.1 (a) depicts the dorsal and ventral pathways in the human brain, while 1.1 (b) demonstrates the large number of connections that link the brain's cortical areas and the two pathways in order to pass on information as suggested by [Felleman and Van Essen, 1991]. The following subsections present the general physiology and an overview of the relevant visual areas.

1.2.1 General physiology

Vision is an intricate process that requires numerous computations generated via a large number of elements in the brain. The core element of the brain is the *neuron* (also known as the nerve cell). The brain contains an incredible 10^{12} cells [Hubel, 1995]. The important morphological parts of the neuron are the *dendrites* and *axon*. The elaborate branching structure of the dendritic tree enables a neuron to receive inputs from other neurons through *synaptic* connections. On the other hand, the *axon* carries the neuronal output to other cells [Dayan and Abbot, 2001]. In essence, a typical neuron has a *soma* (cell body that containing the cell nucleus), *dendrites* (a branched structure that forms the main incoming signals pathway), and a single *axon* (a transmitter of outgoing signals to the surrounding cells).

Neurons propagate pulses or electrical signals in response to a chemical interaction known as the *action potential*. The fundamental mechanism of the action potential was first demonstrated by [Hodgkin, 1937a]; [Hodgkin, 1937b], where the authors blocked the action potential after either cooling or pressuring the nerve segments. The principle of the action potential is based on the differences in voltage (i.e., potentials) between the inside and outside of the cell membrane. The inside of the cell is more negative than the outside (membrane potential is about -70 mv), and the neuron is said to be polarized. The channels of the electrically charged particles or ions (sodium *Na+*, potassium *K+*, chloride *Cl-*, and calcium *Ca2+*) that are located in the neuron membrane are opened or closed in response to neurotransmitters. The redistributions of the ions change the voltage differences across the membrane. As a consequence,

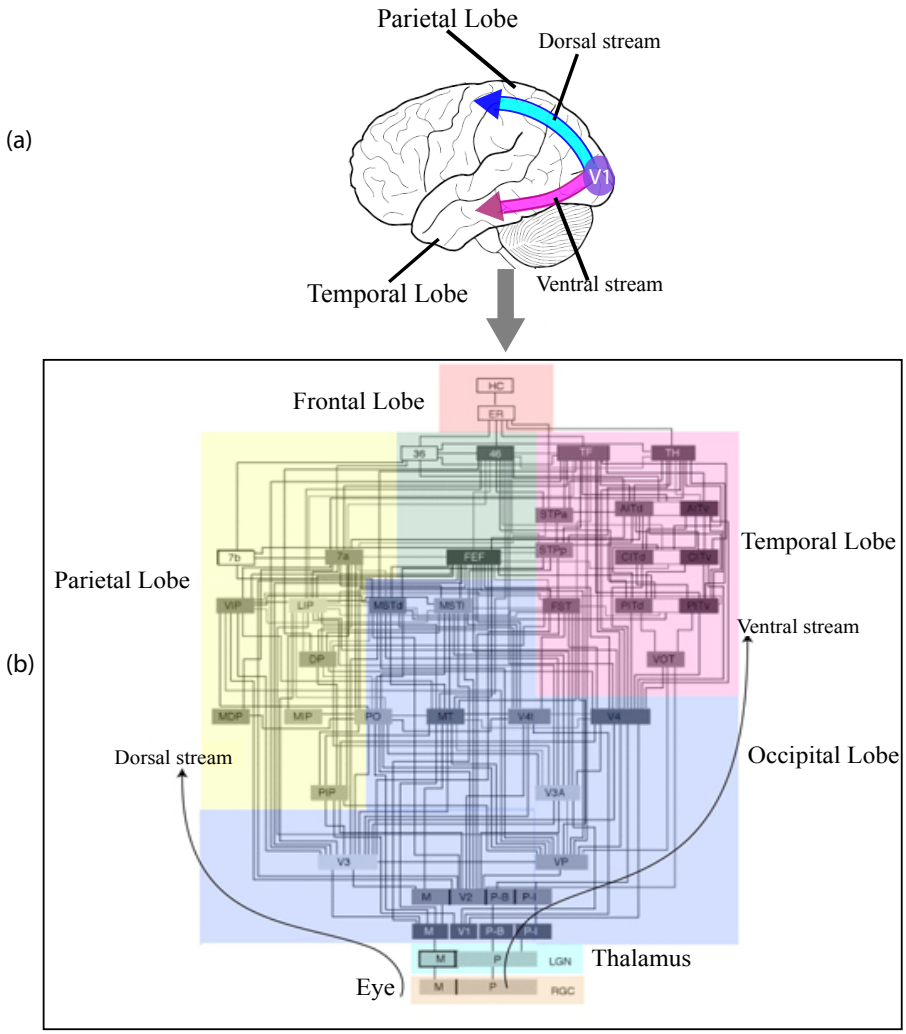


Figure 1.1: The hierarchy of the visual areas. (a) shows a sketch of the human brain in which the main processing pathways are depicted. The dorsal pathway includes areas V1, MT, MST and beyond. This pathway processes the motion information that related to where is the object in the scene. The ventral pathway contains areas V1,V2, V4 and beyond focuses on computing the form information such as extraction the object contour and junctions. (b) illustrates the numerous connections among the hierarchy of the visual areas that suggested by [Felleman and Van Essen, 1991] (The image is adopted from http://vis.berkeley.edu/courses/cs294-10-fa07/wiki/index.php/A1-ArielRokem). The hierarchy shows 32 visual areas that connected by 187 links where most of which have been demonstrated to be reciprocal pathways.

when the voltage difference is decreased (also called depolarized) and reaches a threshold value, an impulse or spike (i.e., action potential) will travel along cell's axon. The production of an action potential is sometimes referred to as *firing*.

Brain cells interact via electrical impulses and are always active all the time so as to be able to manage every day life. Each neuron in the brain receives information from thousands of other nerve cells, and in turn sends information to other thousands of cells. The complexity of the brain is a matter not just of numbers of cells, but the intricacy of brain organization, how the brain works, and the particular properties of each area in the brain. Recent advances in medical imaging technologies (e.g., magnetic resonance imaging (MRI) and Functional MRI (fMRI)) have had a major contribution in understanding the activation in brain areas and to identify neural models for each area (see [Wandell and Dougherty, 2006] for a review of how techniques have been developed for measuring human brain activity). This gives rise to the question regarding the level at which the neural model simulates the functionality of the biological neuron to build an artificial cognitive system. The neural model is a burgeoning topic that requires further investigation. This thesis focuses on the creation of a new neural model for motion detection and integration. The following sections provide an overview on the visual areas considered in this work.

1.2.2 Overview of the visual areas related to this work

- ***The retina*** is the first processing area in the visual system where the photo receptors convert luminance into signals. These signals are passed on via the optic nerve to the next visual area, *the lateral geniculate nucleus* (LGN) . The LGN consists of magnocellular and parvocellular layers in which neurons have a circular receptive field (RF) with a center-surround organization [Kuffler, 1953]; [Hubel and Wiesel, 1961]; [DeAngelis et al., 1995]. Figure 1.2 (A) shows the structure of the RF of the LGN neuron that has an ON (+) and OFF (-) center-surround organization. The ON center responds to bright visual stimuli, while the OFF surround region responds to dark stimuli. The center-surround organization of the neurons' RFs enables these cells to send contrast-information on the visual stimuli based on the differences of the center-surround firing rate [Kandel et al., 2000].

- ***The primary visual cortex*** (also known as V1 or striate cortex) receives the output signals from the LGN. Area V1 is located in the occipital lobe of the brain and plays a critical role in visual perception.

This area is considered as the first processing area of visual information and the basis of two visual processing pathways, the ventral and dorsal pathways [Ungerleider and Mishkin, 1982].

Orientation selectivity is a key feature of the responses of V1 neurons, in which neurons respond strongly to a bar stimulus that is placed at a particular orientation [Hubel and Wiesel, 1959]; [Hubel and Wiesel, 1962]. Another important feature is that the cells in area V1 respond to spatialtemporal orientation (i.e., motion) in which the V1 cells gather inputs from an array of photoreceptors covering the spatial extent of its receptive field, and sum their outputs over time with an appropriate temporal impulse response [Adelson and Bergen, 1985].

In area V1, neurons are classified as simple and complex cells [Hubel and Wiesel, 1960]; [Hubel and Wiesel, 1968]. Neuroscientists have shown that V1 simple cells use linear filters which are polarity-dependent on the bar contrast. The complex cells combine the activity of the simple cells in a nonlinear fashion, and thus the responses of the complex cells are no longer polarity-dependent [Movshon et al., 1978b]; [Movshon et al., 1978a]; [Emerson et al., 1987]; [Pack et al., 2006].

V1 neurons have elongated receptive fields (RFs) due to a single retinal output that is transmitted to area V1 via multiple pathways with different strengths [Kara and Reid, 2003]; [Derrington and Webb, 2004]. By way of illustration, Figure 1.2 shows the receptive field structure of the simple and complex cells in area V1. The simple cells RFs are characterized by separate ON and OFF elongated subregions. The ON (+) subregion responds to the bright visual stimuli, and the OFF (-) subregion responds to dark visual stimuli. In contrast, the complex cells respond to the dark and bright visual stimuli at any position in the RF due to the lack of separate ON/OFF subregions. Along the dorsal pathway, the responses of V1 cells transmit to a subsequent area *middle temporal* (MT) and beyond such as *medial superior temporal* (MST).

- ***The area MT*** is considered as the main processing area for motion perception, in which V1 responses are integrated within larger RFs [Simoncelli and Heeger, 1998]; [Xiao et al., 1997a]; [Xiao et al., 1997b]. This area has been investigated widely (see for example [Zeki, 1974]; [Van Essen et al., 1981]; [Felleman, 1984]; [Born and Bradley, 2005]), and these studies have confirmed that the area MT contains high constructions of direction-selectivity neurons which are highly responsive to visual motion. These studies have identified the responses of area MT as one of the main signals in the dorsal processing stream.

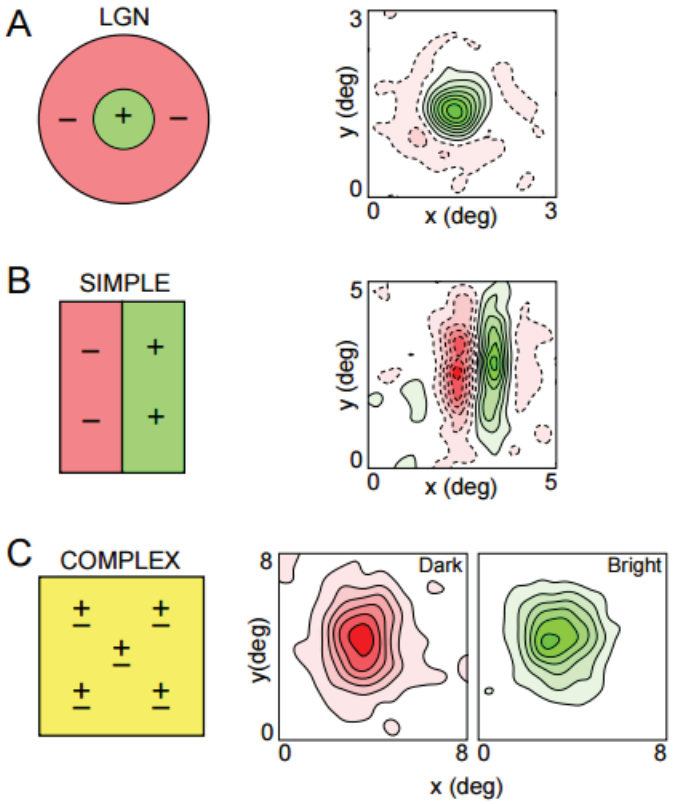


Figure 1.2: Receptive field structure of the neurons in areas LGN and V1 (image taken from [DeAngelis et al., 1995]). (A) Schematic and experimental profiles of the RF of an ON-center neuron from the LGN of a cat. In the traditional depiction (left), the RF has a central "ON" region (+) which is responsive to the onset of a bright stimulus, and a surrounding "OFF" region (-) which is responsive to the onset of a dark stimulus. On the right is shown a 2-D spatial (x-y) RF profile for an ON center cell. Regions of visual space that are responsive to bright spots are delimited by solid contour lines; regions responsive to dark spots are represented by dashed contours. Darkness of shading is proportional to response strength. A center surround structure is clearly seen in this profile, although the surround is fairly weak. (B) The left image schematically depicts the RF of a simple cell exhibits an alternating arrangement of elongated subregions that are responsive to either bright (+) or dark (-) stimuli. A measured RF profile for a simple cell from cat striate cortex is shown on the right as a contour map. (C) Spatial RF structure of a complex cell. In the traditional schematic illustration shown on the left, pluses and minuses are shown throughout the field, indicating that the cell responds to both bright and dark stimuli at each position. Panels on the right show the RF profile of a complex cell. Because regions responsive to bright and dark stimuli overlap, separate profiles are shown for bright and dark stimuli.

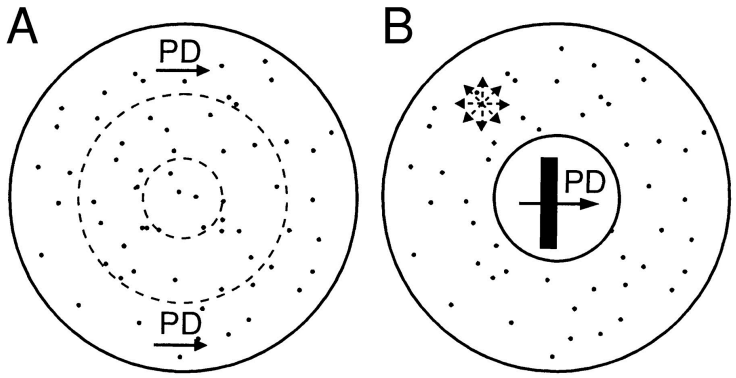


Figure 1.3: Receptive field structure of the MT neurons (image taken from [Born, 2000]). A: classical receptive field (CRF) cell. The neuron was tested with random dot patterns moving within circular windows of varying diameters. The dots always moved in the neuron's preferred direction. B: center-surround cell. The center of the neuron's receptive field was stimulated with a single bar that moved in the preferred direction (PD). The surround was stimulated with an annular window of random dots whose direction of motion varied from trial to trial.

The geometry of MT neurons has been investigated by [Born, 2000], who identified two types of MT cells: center-surround and classical receptive field (CRF) cells, Figure 1.3. The center-surround cells respond well to a single bar or spot but poorly to large textured stimuli such as random dots. This is because receptive fields have antagonistic surrounds, thus rendering neurons unresponsive to wide-field motion. However CRF cells respond best to wide-field motion because neuronal receptive fields have reinforcing surrounds. The nature of the center-surround interactions has also been tested by Born who found that the preferred direction of the surround, with respect to that of the center, tends to be either in the same or opposite direction and rarely in orthogonal directions.

1.2.2.1 Neuron properties in areas V1 and MT

Each area in the visual system has specific properties [Lamme et al., 1998]. Neurons in area V1 are selective for spatio-temporal frequency, while neurons in area MT are more responsive and tend to have a higher directional selectivity than neurons in V1 [Churchland et al., 2005]. V1 and MT neurons have a preferred direction tuning in which V1 cells have a weak inhibition in the anti-preferred direction. The preferred direction of motion is identified as

the direction that elicits the greatest mean firing rate, and the anti-preferred direction is 180° from the preferred direction [Snowden et al., 1991]. Snowden and co-authors showed that the proportion of directionally selective responses is 30% in V1 and 92% in MT, which is consistent with the findings in [De Valois et al., 1982]; [Lagae et al., 1993].

Concerning the speed selectivity of V1 and MT neurons, [Mikami et al., 1986] showed that the large RFs of neurons in area MT permit directional interactions over larger distances than the more limited RFs of V1 neurons. As a consequence, MT neurons have a functional advantage because they detect directional differences for higher speeds than V1 neurons. This is in agreement with the experimental results of [Churchland et al., 2005] and [Priebe et al., 2006], where the authors recorded higher preferred speeds of neuron responses in area MT compared to the responses of neurons in area V1.

The speed tuning of MT neurons has been investigated by [Perrone and Thiele, 2002]; [Perrone and Thiele, 2001]; [Perrone, 2004], who showed that the speed sensitivity of MT neurons is influenced by the degree of orientation of the spectral receptive fields in the spatiotemporal frequency domain. Figure 1.4(a and b) presents a spatiotemporal frequency plot of a moving object at speed v, in which the spectral receptive field lies on a line of slope -v. The slope of the object spectrum changes depending on the object's speed where a faster speed produces a steeper slope. In the spatiotemporal frequency domain shown in Figure 1.4(c), the sampling size of the neuron receptive field grows linearly with the distance from the origin in order to keep the same uncertainty in the velocity space [Grzywacz and Yuille, 1990].

1.2.2.2 Motion perception and aperture problem

Individual neurons, in principle, can respond either to the motion of each orientated contour (component motion) or to the overall movement of the pattern (pattern motion) [Snowden et al., 1991]. This has been investigated by [Adelson and Movshon, 1982]; [Movshon et al., 1985] where the authors showed that all neurons in area V1 respond in component motion, while numerous numbers of MT neurons respond in a pattern motion (see also [Rodman and Albright, 1989] for discussions regarding the underlying motion perception of neurons). Movshon and co-authors have attributed this to the existence of two stages of motion analysis in the mammalian visual system. The first stage deals with motion information perpendicular to the contrast orientation, which is then integrated into the second stage based on a nonlinear fashion.

Neurons in the primary visual cortex (V1) are selective to spatio-temporal stimulus features where neurons in this area have small RFs, or filter sizes. Consequently, they can only detect local motion components that occur within

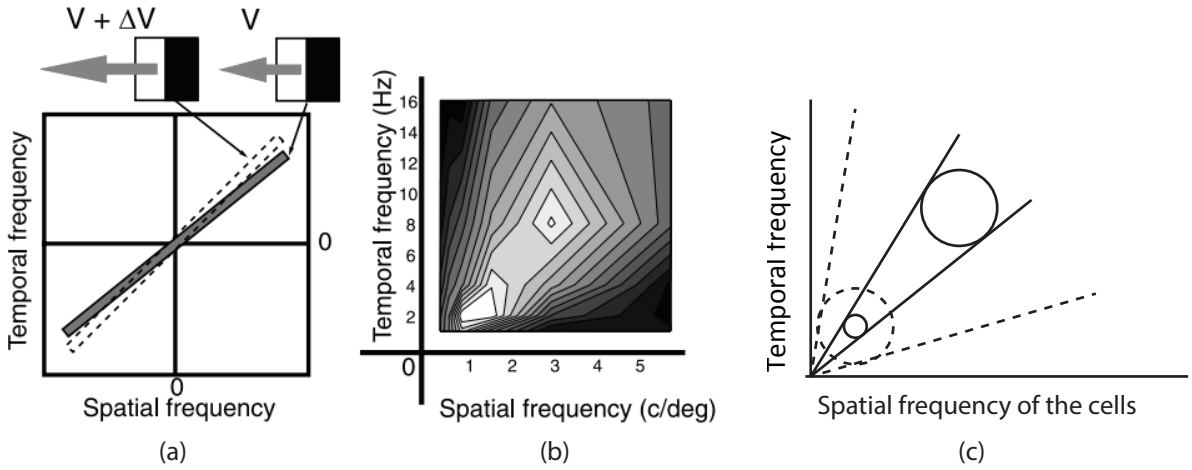


Figure 1.4: Spatiotemporal frequency plots. (a) Representation of moving edge spectra for two edges, one moving right to left at speed v (shaded outline) and the other at v+Δv (dashed outline). Discrimination of different edge speeds requires a mechanism that can respond selectively to a particular slope of the edge spectrum. (b) Spectral receptive field (SRF) of an MT neuron tested with 30 different combinations of spatial and temporal frequency (Figures (a) and (b) taken from [Perrone and Thiele, 2002]). (c) The sampling size of the neurons, the diameter of the circles represents the frequency bandwidth of the cell. The angle between the lines starting at the origin and tangential to the circles represents the velocity uncertainty of the cells. If all the cells have the same bandwidth, cells near the origin would have large uncertainty (angle between dashed lines). on the other hand, if the bandwidth scaled with distance from the origin, then all cells would have the same uncertainty (angle between solid lines) (the image adopted from [Grzywacz and Yuille, 1990].

their RFs. This means that along elongated contrasts only ambiguous motion information can be detected locally. It is the normal flow component that can be measured along the local contrast gradient of the luminance function (aperture problem).

The aperture problem can be resolved in area MT by integrating several normal flows that are estimated at distant locations in area V1. The integration strategy can be either based on vector integration (VA) [Yo and Wilson, 1992] or on the intersection-of-constraints (IOC) mechanism, as suggested by [Adelson and Movshon, 1982]. The latter approach can be demonstrated to calculate the exact movement of, for example, two overlapping gratings which move translatory behind a circular aperture in distinct directions (plaid). If the two gratings have the same contrast and spatial frequency, the plaid appears as a single pattern that moves in the direction of the intersecting normal flow con-

straint lines defined by the component gratings. This direction corresponds to the feature motions generated by the grating intersections. Another approach is to integrate local motion responses of unknown components and compositions using a biologically-inspired motion integration as reported in [Bayerl and Neumann, 2004] and [Bouecke et al., 2011]. The authors used model mechanisms of in the cortical area MT that integrate initial V1 cell responses. The RFs of cells in MT are larger in their size by up to an order of magnitude. In other words, such cells operate at a much larger spatial context to properly integrate localized responses, similarly to the VA method. In addition, the authors of [Bayerl and Neumann, 2004] and [Bouecke et al., 2011] considered differentially scaled responses generated by the output normalization in V1. As a consequence, localized feature responses at line ends or corners lead to stronger responses in the integration process. This thus leads to a hybrid mechanism of weighted mixed vector integration and feature tracking.

1.3 Neural modeling

The core processing unit in neural modeling is individual neurons. The dynamical model of individual neurons can be described at different abstraction levels. In [Herz et al., 2006] the authors reviewed the principles of single-cell modeling at five main levels, see Figure 1.5. *Detailed compartmental models* represent the first main level of single-cell modeling which closely resembles the biophysical-mathematical models of the neurons. These models explicitly consider all the spatial structure and potential complexities of the neuron [Bower and Beeman, 1998]. The theoretical basis of these models is the *cable theory* provided by Rall's studies (see [Segev et al., 1995]). Although detailed compartmental models can approximate the dynamics of single neurons quite well, their high dimensionality and intricate structure rule out any mathematical understanding of their emergent properties. As a consequence, these models are not well suited for large-scale network simulations [Herz et al., 2006]. *Reduced compartmental models* overcome these problems by retaining the general morphology of the pyramidal cells (somato-dendritic interaction) using only one or a few dendritic compartments. The reduced models run significantly faster than the full models but they are not suitable for studying the effect of single synaptic inputs on single dendritic branches of pyramidal cells or local dendritic processing in general (see [Bush and Sejnowski, 1993]; [Kuznetsov et al., 2006]; [Herz et al., 2006]). *Single-compartment models* disregard the neurons spatial structure and consider it as a point process by highlighting how spikes or action potentials are generated. Such models are presented in [Hodgkin and Huxley, 1952a]; [Hodgkin and Huxley, 1952b]. These models have led to

a quantitative understanding of many dynamical phenomena including phasic spiking, bursting, and spike-frequency adaptation (see [Koch, 1998]; [Herz et al., 2006]). The fourth type of individual neuron models are *cascade models* in which a single neuron is modeled on a more conceptual level based on a concatenation of mathematical primitives, such as linear filters, nonlinear transformations, and random processes. The fifth type of individual neuron model is a *black-box* where signal processing of the individual neuron is considered, while biophysical machinery is ignored.

In this thesis we used a *single-compartment model* in which individual neuron is represented as a like-point process, ignoring the influences of the complex dendritic trees and the relative nonlinear interactions. This kind of representation helps in understanding many dynamic phenomena and can be found in several models (e.g. [Bouecke et al., 2011]; [Bayerl and Neumann, 2004]; [Tschechne et al., 2014a]).

The Neuron's membrane plays a key role in the neurons where it acts as a resistor to prevent various ions from passing freely via the barrier and as a capacitance to construct a charge inside and outside of the cell. Gates of the membrane permit various quantities of ions to pass across a barrier in order to alter the potential on both sides of the membrane. The cell is said to be in a dynamic equilibrium state if there is no input current (zero net current flow). The basic dynamic description of the neuron's membrane can be defined in the RC circuit in which the potential (voltage) dynamics can be determined using Kirchhoffs law, namely,

$$R \cdot C \cdot \frac{dv\,(t)}{dt} = -v\,(t) + R \cdot g_{ex} \cdot (E_{ex} - v\,(t)) + R \cdot g_{in} \cdot (v\,(t) - E_{in})\,, \quad (1.1)$$

where R represents the resistance, C denotes the conductance, v defines the resting potential, E_{ex} and E_{in} define the saturation points of the reversal battery potentials, and g_{ex} and g_{in} represent excitatory and inhibitory synaptic inputs (these inputs are received via fast chemical synapses). The cell generates a spike or pulse when the synaptic inputs trigger a depolarization that exceeds a threshold value where this process is known as action potential and has been summarized by [Grossberg, 1988]. Grossberg defined the feedforward neuronal responses using a generalized concept of a membrane equation

$$\tau \frac{dv\,(t)}{dt} = -A \cdot v\,(t) + (B - C \cdot v\,(t)) \cdot net_{ex} - (D + E \cdot v\,(t)) \cdot net_{in}, \quad (1.2)$$

where A represents the passive leakage, B and D denote the saturation

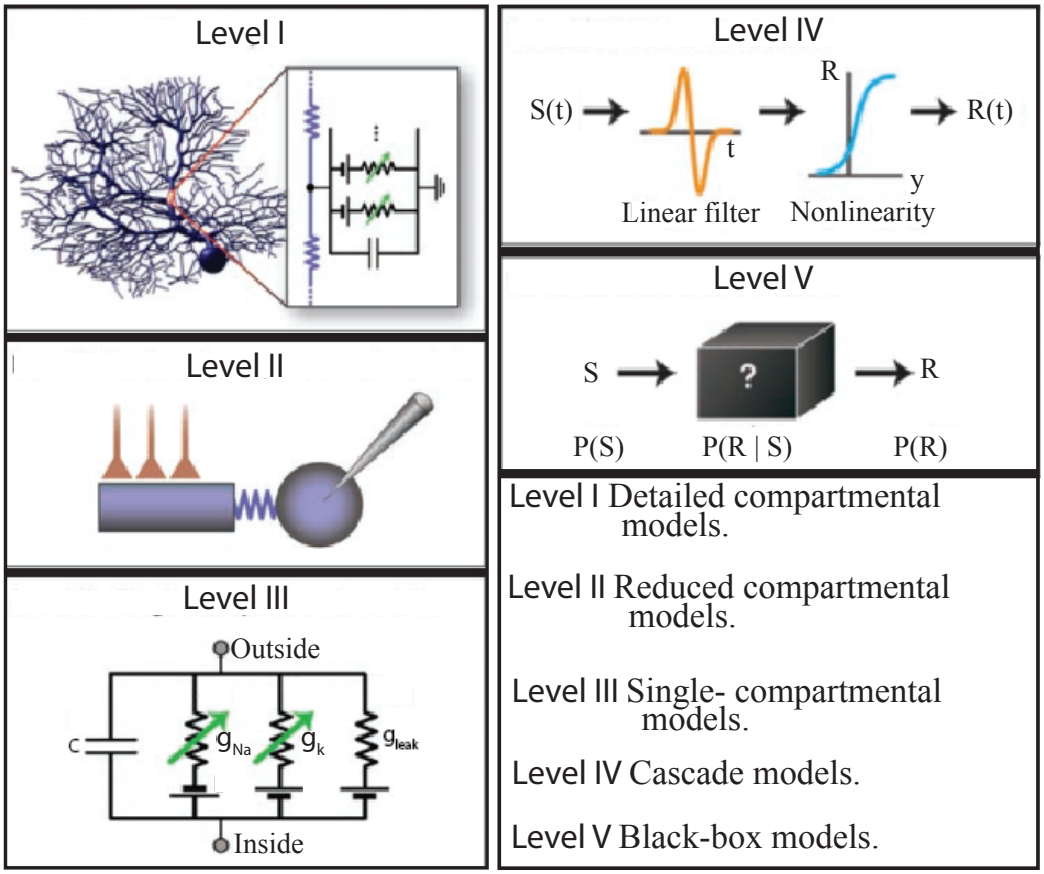


Figure 1.5: Examples for five levels of single-cell modeling. Level I: Detailed compartmental model. The dendritic tree is segmented into electrically coupled Hodgkin-Huxleytype compartments (level III). Level II: Two-compartment model. The dendrite receives synaptic inputs and is coupled to the soma where the neurons response is generated. Level III: Hodgkin Huxley model, the prototype of single-compartment models. The cells inside and outside are separated by a capacitance C and ionic conductances in series with batteries describing ionic reversal potentials. Sodium and potassium conductances (g_{Na}, g_K) depend on voltage; the leak g_{leak} is fixed. Level IV: Linear-nonlinear cascade. Stimuli S(t) are convolved with a filter and then fed through a nonlinearity to generate responses R(t), typically time-dependent firing rates. Level V: Black-box model. Neglecting biophysical mechanisms, conditional probabilities p(R—S) describe responses R for given stimuli S (the image is adopted from [Herz et al., 2006]).

potential parameters (relative to C and E), and net_{ex} and net_{in} denote generic excitatory and inhibitory inputs to the target cell. Equation 1.2 represents the basis of the notional format used in this thesis. The saturation properties were investigated at equilibrium state where the state response is calculated by $\frac{dv(t)}{dt} = 0$. The resulting $v\,(t)$ response is finally calculated by

$$v\,(t) = \frac{B \cdot net_{ex} - D \cdot net_{in}}{A + C \cdot net_{ex} + C \cdot net_{in}}. \tag{1.3}$$

The upper and lower bounds of v_t can be determined when the limitation of the excitatory and inhibitory inputs net_{ex} and net_{in} approaching ∞, i.e, $\lim_{net_{ex} \to \infty} v\,(t) \cong B/C$ (upper bound) while $\lim_{net_{in} \to \infty} v\,(t) \cong -D/E$ (lower bound).

1.4 Motion detection related techniques

Motion detection has been widely investigated in several fields such as robot vision, tracking, computer vision, surveillance (e.g., [Moeslund and Granum, 2001]; [Murray and Basu, 1994]; [Green et al., 2004]). The main techniques for motion detection fall into three categories ([Barron et al., 1994]; [Simoncelli, 1993]; [Escobar, 2009]):

(i) Gradient techniques: Also known as *differential* techniques in which the optic flow is estimated based on the derivatives of image intensity over space and time.

(ii) Frequency-based techniques: These techniques use *spatiotemporal oriented filters*, where the motion is derived in the Fourier domain (Frequency space).

(iii) Correlation-based or matching techniques: These techniques are based on matching particular *features* in small regions of the subsequent images. The matching criteria is usually the least squares method or normalized correlation measure.

We focused on the frequency-based techniques (see subsections bellow) and the gradient techniques since the matching techniques are not a biologically oriented approach.

1.4.1 Gradient techniques

The gradient techniques use spatial and temporal partial gradients to compute the local optic flow in an image. These techniques operate under the assumption of a *brightness constancy constraint*, where the intensity of a moving scene does not change. This means that the intensity of a moving object along two frames that are captured at time t and Δt remains constant. Thus, the image brightness constancy equation can be defined by

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t), \tag{1.4}$$

where I(x,y,t) represents the intensity at location (x,y) and time t, whilst $I(x + \Delta x, y + \Delta y, t + \Delta t)$ denotes the intensity after a spatial displacement $(\Delta x, \Delta y)$ along x,y directions at time $(t + \Delta t)$. According to the assumption of brightness constancy, the derivative of the image intensity function I(x,y,t) with respect to time t is equal to zero ($\frac{dI(x,y,t)}{dt} = 0$). The constancy equation is thus given as

$$\frac{\partial I}{\partial x} v_x + \frac{\partial I}{\partial y} v_y + \frac{\partial I}{\partial t} = 0, \tag{1.5}$$

where $v_x = \frac{\partial x}{\partial t}, v_y = \frac{\partial y}{\partial t}$ are the velocity components along x and y directions, which are also known as the optic flow vector $\mathbf{v} = (v_x, v_y)^T$ (for details see [Fleet and Weiss, 2006]). Equation 1.6 can be written as

$$\bigtriangledown I(x, y, t) \cdot \mathbf{v} = -I_t(x, y, t), \tag{1.6}$$

where $I_t(x, y, t)$ represents the partial derivative of $I(x, y, t)$ over the time axis, $\bigtriangledown I(x, y, t)$ defines the partial derivative $(I_x(x, y, t), I_y(x, y, t))^T$ over spatial position (x, y). Equation 1.6 cannot be solved as such because it is one equation and has two unknowns parameters ($\mathbf{v}$). This is known as the *aperture problem* of the optical flow algorithms in which the intensity gradient constrains the flow to one parameter family of velocities along a line in velocity space [Fleet and Jepson, 1990]. This line is perpendicular to $\bigtriangledown I(x, y, t)$, and its perpendicular distance from the origin is $|I_t(x, y, t)| \,/\, \|\bigtriangledown I(x, y, t)\|$.

In order to solve Equation 1.6, an additional constraint should be considered. Several optic flow algorithms have applied additional constraints to solve this equation. We review some of them below.

In [Nagel, 1987, Tretiak and Pastor, 1984], the authors suggested using second order derivatives of I to constrain 2D velocity. The second order derivative can be calculated based on the constancy of $\bigtriangledown I(x, y, t)$ which leads to $\bigtriangledown I(x, y, t)/dt = 0$. This is also known as *Hessian* of I which can be defined by

$$\begin{bmatrix} I_{xx}(x,y,t) & I_{yx}(x,y,t) \\ I_{xy}(x,y,t) & I_{yy}(x,y,t) \end{bmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} + \begin{pmatrix} I_{tx} \\ I_{ty} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{1.7}$$

This method has a stronger limitation than Equation 1.6 due to the conservation of $\bigtriangledown I(x,y,t)$ which implies that first order deformations of intensity (e.g., rotation or dilation) should not be present [Barron et al., 1994]. In [Girosi et al., 1989], the authors suggested using the constraints of Equation 1.7 in addition to the constraints of Equation 1.6. Since the second order derivative cannot usually be measured accurately enough to determine the tangential component of $\mathbf{v}$, optic flow estimations from second order method are sparser and less accurate than the estimation by first order methods [Barron et al., 1994].

The assumption of the nearby pixels with the same 2D velocity is another method to constrain $\mathbf{v}$ (i.e., $\mathbf{v}$ is taken to be constant over a small spatial neighborhood). Here, the least squares technique is used to minimize the constraint errors. The least squares error estimator can be defined as

$$E(\mathbf{v}) = \sum g(x)[\bigtriangledown I(x,y,t) \cdot \mathbf{v} + I_t(x,y,t)]^2, \tag{1.8}$$

where g(x) represents a weighting function of the spatial neighborhood. Usually, the Gaussian function is used as a weighting function in which the center constraints have a higher weight than the periphery constraints. Although this method gives a good estimation [Lucas and Kanade, 1981], it is estimated locally and has no global notion.

1.4.2 Frequency-based techniques

Frequency-based techniques estimate the optic flow using spatio-temporal filters tuning in the Fourier domain. Compared with other approaches, using spatiotemporally oriented energy in Fourier space makes these techniques more capable of estimating motions, such as random dot patterns [Adelson and Bergen, 1985, Beauchemin and Barron, 1995].

Frequency-based filters can generally be classified as *energy-based* and *phase-based*. Energy-based filters use the energy responses of the spatiotemporal filters to compute the motion (e.g. [Adelson and Bergen, 1985]), while the phase-based filters detect image velocities based on the local image phase (see e.g. [Fleet and Jepson, 1990]).

This section provides an overview of the principles of spatiotemporal filters and focuses on the mechanism suggested by [Adelson and Bergen, 1985] and adopted in our work. Spatiotemporal filters can be modeled using Gabor filters which provide a close description of the receptive fields in the primary visual

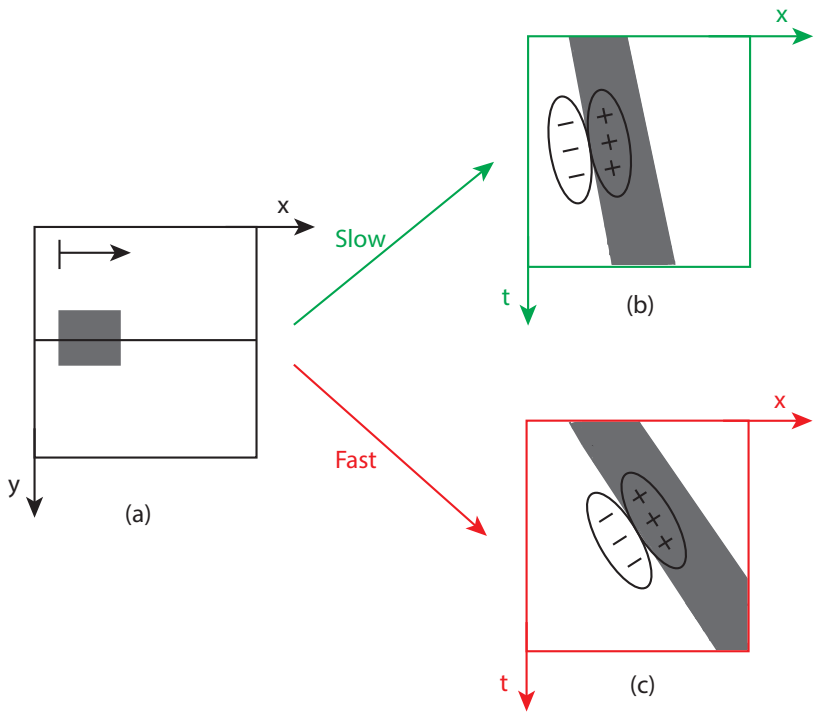


Figure 1.6: Motion representation of a black box. (a) the box is moving from left to right in 2D space (x,y) at two speeds. (b) box is moving with slow speed (green frame). (c) box is moving with fast speed (red frame). Box movement traces a slanted bar in the spatiotemporal domain in which higher speed generates steeper bar respect to t-axis in the 2D space-time (x,t) domain (motion representation in (y,t) is ignored). It also shows the two lobes (positive and negative) of the spatiotemporal filter with an orientation that matches to the orientation of the bar.

cortex, area (V1) [Ringach, 2002]. These filters are oriented in the space-time domain (x, y, t). The orientation in the space-time domain construes motions as an orientation in the spatial and time domain (x, y, t). Figure 1.6 (a) shows a box that moves from left to right with different speeds. In 2D space (x,t), Figure 1.6 (b,c), the motions of the box drive slanted bars corresponding to the speed of the box (fast motion drives more slanted bars with respect to the t-axis than slow motion). The motion can be detected when the oriented filter matches the movement of the input stimulus in the spatiotemporal domain. The idealized matching thus requires a set of oriented filters, where each one responds best to the motion that coincides with their orientation. Consequently, spatiotemporal filter mechanisms depend on the orientation of the motion in space-time domain.

In addition, the responses of the spatiotemporal filters rely on the contrast of the input stimulus (e.g., for a moving sinusoidal grating, the responses of the filters oscillate sinusoidally over time). To tackle the problem of polarity-dependence, [Adelson and Bergen, 1985] suggested *motion energy* filters (see the shaded block in Figure 1.7) which combine the responses of two spatial filters with different phases. The spatial filters can be constructed using an odd and even component of 2D spatial Gabor filters, or by simply shifting the filters spatially by 90°. The common definition of a Gabor filter is given by [Daugman, 1985]

$$\begin{aligned} G(x, y, \sigma_x, \sigma_y, \xi_x, \xi_y) &= exp(-\pi[(x - x_o)^2\sigma_x^2 + (y - y_o)^2\sigma_y^2]) \cdot \\ & \quad exp(-2\pi i[\xi_x(x - x_o) + \xi_y(y - y_o)]), \end{aligned} \tag{1.9}$$

where σ_x and σ_y control the width of the spatial extent along the x and y axes, respectively. ξ_x, ξ_y are the frequency tuning parameters of the 2D Gabor function $G(\cdot)$. This type of 2D elementary function constructs a generalization of the 1D elementary functions originally suggested by [Gabor, 1946]. Shifting and similarity of Gabour functions are reflected in the position parameters (x,y), while the modulation and scaling of the functions are reflected in the modulation parameters ξ_x, ξ_y and two scale parameters σ_x and σ_y (for details see [Daugman, 1988]).

Regarding the temporal filters, Adelson and Bergen suggested using gamma functions of different durations in order to achieve temporal smoothing and differentiation, thus leading to a temporally biphasic response shape. The temporal filters responses represent the time course of the neurons, which highlight how past inputs are weighted and integrated to produce the response at that moment. The temporal filter is defined by

$$f(t) = (kt)^2 \cdot exp(-kt^2) \cdot [1/n! - (kt)^2/(n + 2)!] \tag{1.10}$$

Figure 1.7 reviews all the stages of the motion energy model suggested by [Adelson and Bergen, 1985]. In Adelson and Bergen's scheme, the input signal I(x, t) passes through the spatial and temporal filters to produce four separable responses A, A', B, and B'. Sums and differences of these responses produce spatiotemporally oriented linear responses, (A-B', A'+B, A+B', A'-B). The resulting responses are selective for motion direction, leftward and rightward. The oriented linear responses are characterized as phase-dependent responses in which the response to a light bar has the opposite sign of the response to a dark bar. Response pairs for leftward and for rightward motions are combined by summing their squares. The non-linear combination drives

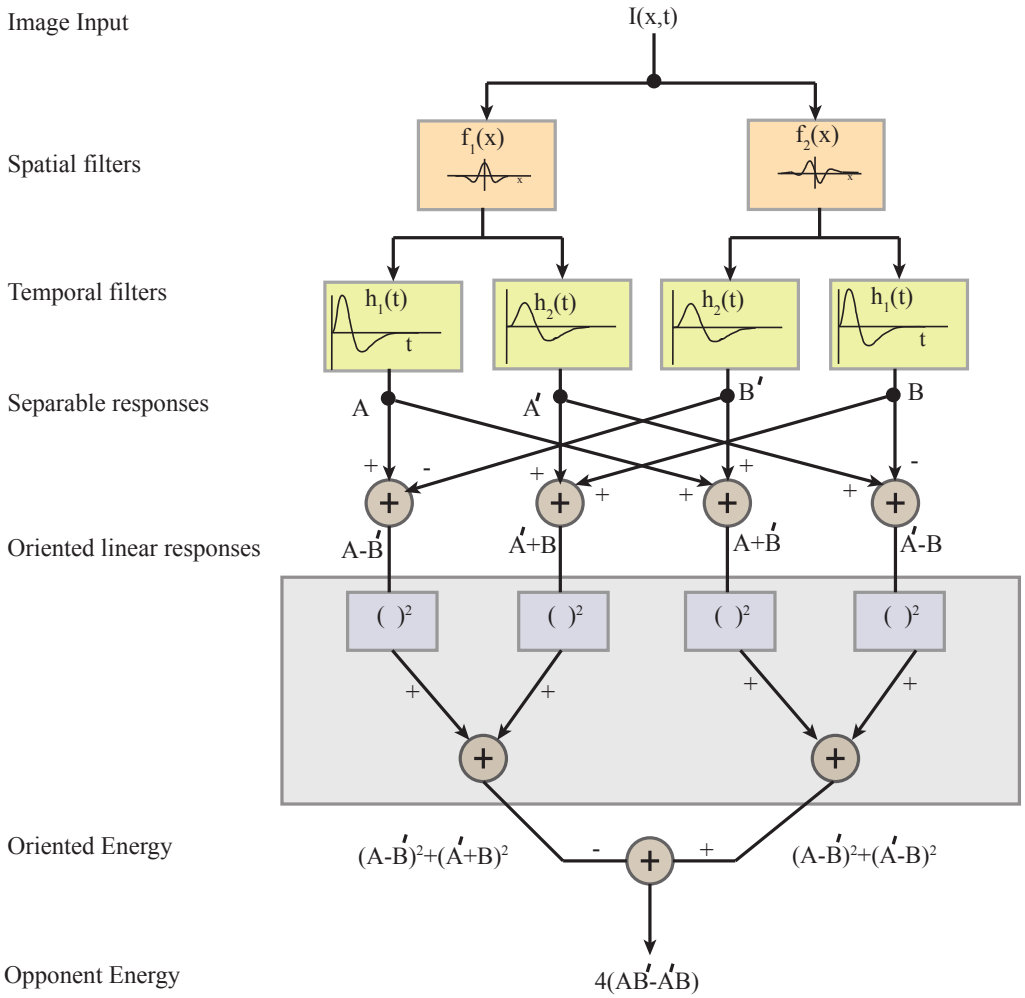


Figure 1.7: Motion energy model suggested by [Adelson and Bergen, 1985]. The visual input I(x, t) passes through the two spatial filters with different phases (whose responses 90° out of phase) f1(x) and f2(x). Each output passes through two biphasic temporal filters with different duration h1(t) and h2(t) to generate four separable responses, A, A', B, and B'. These outputs are summed and subtracted to generate the oriented linear responses. Squares of quadrature pairs of the oriented linear responses are summed to generate the motion energy of each direction (oriented energy). The difference between the rightward and leftward signals gives the opponent energy.

the oriented-energy responses, $(A-B')^2+(A'+B)^2$ and $(A+B')^2+(A'-B)^2$. Here, the responses are no longer phase-dependent in which similar responses are generated for white-black and black-white edges that are moving in the same direction. Therefore, filters in this stage are insensitive to the polarity of the stimulus contrast, but are sensitive to the direction of motion. The difference between the oriented-energy responses generates opponent energy responses. In this stage, the sign of the responses reflects the direction of the motion and is independent of the polarity of the input stimulus. By using simple algebra, the final output reads $4(AB'-A'B)$. This output is the same as the output of the Reichardt type model introduced in [Reichardt, 1961], except for the scale factor. In the Reichardt model, the opponent motion is computed inherently without producing oriented linear responses to the leftward and rightward motion.

Spatiotemporal filters have been widely used in the literature (e.g., ([Hubel and Wiesel, 1962]; [Poggio and Reichardt, 1973]; [Fahle and Poggio, 1981]; [Fleet and Jepson, 1984]; [Van Santen and Sperling, 1984]; [Adelson and Bergen, 1985]; [Heeger, 1988]; [Grzywacz and Yuille, 1990]; [Conway and Livingstone, 2003]) where these models are physiologically motivated. They are distinct from the other techniques because they are biologically-based computational models and closely resemble the functionality of neurons in area V1. The principles of spatiotemporal filters were adopted in this thesis, in which the temporal filters of the type suggested by Adelson and Bergen were adapted to be consistent with a new vision-sensor technology, namely an event-based sensor. This is introduced in detail in Section 2.5.2.

1.5 Outline of the thesis

The following chapters present the details of our proposed neural model as well as the computational algorithm. Each chapter contains an introduction, particular model algorithms and results along with the corresponding discussion.

Chapter 2: We propose an approach for motion estimation using a neuromorphic vision sensor. The model mechanism is inspired by the motion pathway of the visual system and is consistent with the vision sensor functionality. We employ the functionality of the Adelson-Bergen spatio-temporal filtering to create a representation of the input stimulus. The input is generated by an event-based sensor that is sensitive to temporal changes in the input luminance (Plenoptic function). The principles of event-based sensors are demonstrated in Section 2.2. The model extracts motion information via a spatiotemporal energy mechanism, which is oriented in the space-time domain

and tuned in spatial frequency. We also introduce a normalization process to balance activities of individual cells against the neighborhood activities. We tested our model using different kinds of stimuli moved via translatory and rotatory motions. We also show the robustness of our model by probing it with synthetically generated ground truth stimuli and realistic complex motions, e.g. biological motions and a bouncing ball.

Chapter 3: We extend our neural model by incorporating a feedforward and feedback interaction mechanism. In this chapter, we derive a deeper insight about how the visual cortical areas interact and communicate with each other based on different representations. We show how the responses of area V1 are weighted and pooled by area MT cells which are selective to different velocities, i.e. direction and speed. Such feature selectivity is derived here from compositions of activities in the spatio-temporal domain and integrating over larger space-time regions (receptive fields). We suggest a mechanism by which the feature selectivity in both areas can be matched. To achieve such a linkage we integrate the responses over different speeds along a particular preferred direction. The normalization of activities is carried out over the spatial as well as the feature domains to balance the activities of individual neurons in model areas V1 and MT. Our model is tested using different stimuli that move in different directions.

Chapter 4: We show the usability of our neural model for classification purposes. We select action recognition as a case study. We utilize different types of biological motions that are used in our daily life. We use a boundary window to focus on the action where this window is located in the center of human movements. Motion dynamics of the human body and limbs are extracted using spatial sampling regions. We use the support vector machine (SVM) classifiers for action recognition in which an all pairs SVM strategy is used for action classification. We demonstrate the effect of the extracted features on the recognition rate. In addition, we show how the selected subjects in the training set have an impact on the classification rate.

Chapter 5: We summarize the main contributions of this thesis.

Chapter 2

Motion Detection Using Address-Event Representation

2.1 Introduction

Visual processing methods to interpret motions in real-world scenarios present significant challenges. The term *frame-based* (integration at a fix temporal rate) for data acquisition has been in common usage for many years. A frame-based camera transmits moving scenes into a series of consecutive frames. These frames are constructed at a fixed time rate, which generates an enormous amount of redundant information. In other words, conventional frame-based cameras capture visual scenes at a fixed sampling rate (about 30 Hz) in which the intensity of a single position might be change over several frames. As a consequence, consecutive frames contain more information than is required for motion analysis thus entailing a high memory capacity and considerable computational costs. Moreover, conventional frame-based cameras do not provide a high temporal resolution that can cope with the fast motion in which local intensity changes are integrated over a fix sampling rate (standardly $\cong$ 33.3 ms). As a result, fast motion such as flickering or the rapid movement of an object cannot be differentiated.

Many studies have been conducted to improve the properties of cameras. Although resolution, sampling rate and storage capacity have all been enhanced, the concept of frame-based remains unchanged. This concept is far from the principles of the biological visual system where the biological retina generates signals asynchronously that correspond directly to temporal variations in the intensity of the input signals ([Shepherd, 1990]; [Roska et al., 2006]; [Roska and Werblin, 2003]). Recently, *neuromorphic vision sensors* (silicon retinas) have created frame-free visual data acquisition in which this kind

of sensor mimics the neural circuitry of the biological retina.

The first silicon retina or non-frame-based vision sensors were introduced by [Mahowald, 1994] whose research actually started in 1991 (see [Mahowald and Mead, 1991]; [Mahowald, 1992]). The authors demonstrated the functionality of major retina neurons and showed the potential of biological neural simulation. Recently, neuromorphic designers [Lichtsteiner et al., 2008] have constructed silicon retinas (Dynamic Vision Sensors-DVSs) that simulate the transmission of biological retinal responses. The silicon retina transmits outputs as *Address-Event Representation* (AER) spikes [Liu and Delbruck, 2010]. This thesis makes use of DVSs and the next section details how these sensors work.

2.2 Dynamic vision sensors

Dynamic Vision Sensors (DVSs) were introduced in a pioneering work of the (iniLabs) Neuroinformatics Institute at the University of Zurich and the ETH Zurich. These sensors model the circuitry of the biological retina at the pixel-level using an address-event representation [Delbruck et al., 2010]. Each pixel in the sensor implements specialized processing functions that mimic the functionality of individual photoreceptor, horizontal, biopolar and ganglion neurons in the retina (see Figure 2.1). The DVS sensor uses an *Address-Event Representation* (AER) principle with a resolution of 128×128 pixels [Delbruck and Lichtsteiner, 2008]. Neuromorphic circuits simulate the transient responses of the biological retinas using analog electronic circuits (CMOS vision sensors) in which the events appear in the output of the sensor as a polarity ON or OFF format. These events signify scene reflectance changes [Lichtsteiner et al., 2008].

Silicon retinas have a high temporal resolution, where the events are generated asynchronously and sent out almost instantaneously on the address bus, thus, subtle and fast motions can be detected. In addition, these cameras have a low latency and a large dynamic range due to the pixels locally responding to relative changes in intensity. A DVS's ability to produce an event at 1 μs time precision and a latency of $15\mu s$ with bright illumination are illustrated in [Lichtsteiner et al., 2008]. From an engineering perspective, silicon retinas have a higher dynamic range in illumination (> 105) and higher effective sampling rates (several kHz) over conventional cameras, which have a typical dynamic range of a factor of 300 in illumination and sampling rates of under $60Hz$ [Liu and Delbruck, 2010]. The cost of processing the sparse output of silicon retinas can be reduced by a factor of 100 compared to the cost of processing the outputs of conventional cameras ([Delbruck et al., 2010]; [Liu and

Delbruck, 2010]).

Recently, [Brandli et al., 2012] presented *a dynamic and active pixel vision sensor* (DAVIS) that simultaneously combines the output of asynchronous DVS events and synchronous global shutter frames. The combination of the frame-based image readout and asynchronous event-based temporal contrast readout enables the DAVIS to capture information on static as well as dynamic visual scene contents. The spatial resolution of the DAVIS is 240×180 where these sensors have also been provided by iniLabs.

2.2.1 Address-Event Representation (AER)

Asynchronous neuromorphic DVSs generate an output in the form of *address-events*(AEs). Each pixel in the DVS camera generates an event independently and asynchronously when a local change in intensity is detected. Such events are indexed using the 2D spatial location (x,y), and the time-stamp of the event t. The DVS uses the polarity format for event representation, ON (+1) or OFF (-1) if a positive or negative change in the log of intensity is detected. On the other hand, when there are no changes in intensity, no reaction from the sensor is generated (i.e., invariant intensity does not trigger the sensor to generate a response), and as a consequence, any such redundant information sampled by frame-based cameras is reduced. This principle can be defined by

$$e(x,y,t) = \begin{Bmatrix} ON(+1) & if & \frac{d}{dt}(log(I(x,y,t)) > th_{on} \\ OFF(-1) & if & \frac{d}{dt}(log(I(x,y,t)) > th_{off} \\ no\ response & & otherwise \end{Bmatrix}, \tag{2.1}$$

where $e(x,y,t)$ represents the event generated at position (x,y), and time-stamp t. th_{on} and th_{off} denote the ON and OFF threshold values, respectively. In order to show how the ON and OFF events are generated in response to an input stimulus, Figure 2.2 presents an example of the DVS output for a *black-cross stimulus*. This stimulus was rotated clockwise as shown in Figure 2.2(a_1). Since the DVS camera generates an event when changes in intensity are detected, the events are only generated along the stimulus edges. We repeated the recording using a similar stimulus with gray-level smoothing in the interior, namely *smooth-cross stimulus*, as shown in Figure 2.2(b_1). Here the events are generated over the interior as well as the edges of the stimulus, due luminance changes over the whole stimulus (body and edge), see Figure 2.2(b_2).

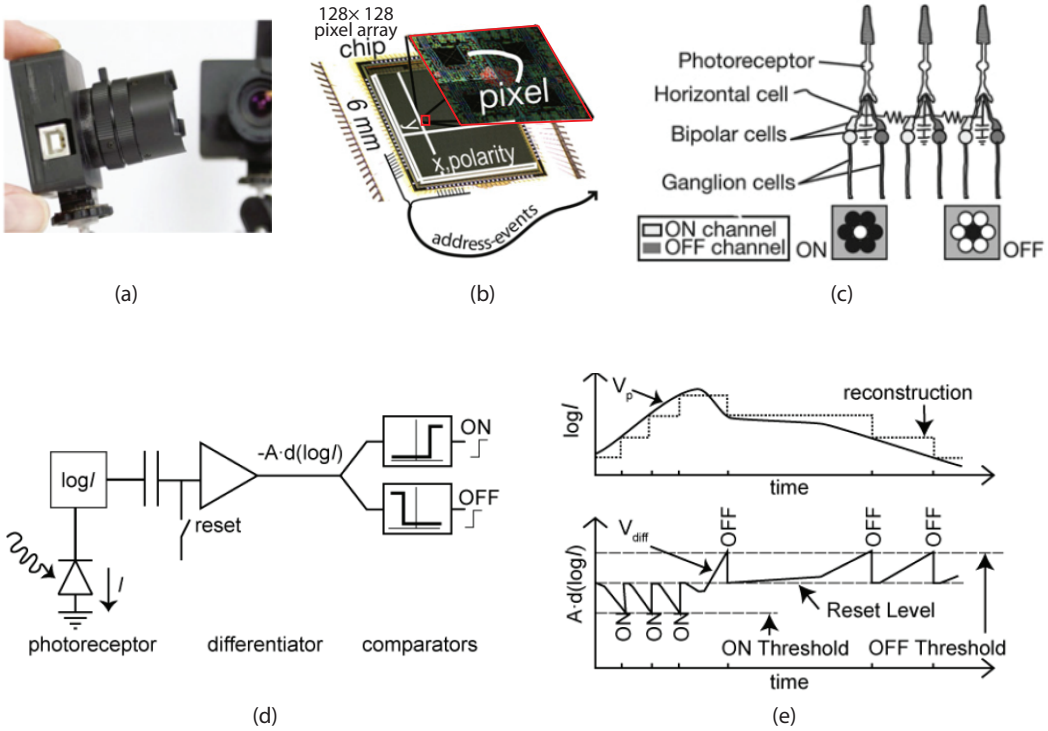


Figure 2.1: The Dynamic Vision Sensor silicon retina. (a) The dynamic vision sensor with lens and USB interface. The camera is interfaced either to a computer by USB, directly to a microcontroller, or to another neuromorphic chip via its AER interface. (b) Chip micrograph and 4-pixel layout. The pixels are arranged in an array and fabricated in a standard CMOS process. Address-Event Representation (AER) circuits along the periphery of the chip handle the access to the shared AER bus and ensure that all events are transmitted, even if there are collisions. Colliding pixels must wait their turn for access to the AER bus. (c) The DVS pixel models a simplified 3-layer retina and forms an abstraction of the photoreceptor-bipolar-ganglion cell information flow. (d) The DVS pixel consists of three parts: a logarithmic photoreceptor, a differencing amplifier (bipolar cells), and two decision units (ganglion cells). The pixel output consists of asynchronous ON and OFF address-events that signal scene reflectance changes. (e) The spike events are computed by the pixel as illustrated. The continuous-time photoreceptor output, which encodes intensity logarithmically, is constantly monitored for changes since the last event was emitted by the pixel. A detected change in log intensity which exceeds a threshold value results in the emission of an ON or OFF event. The threshold is typically set to about 10% *contrast. Communication of the event to the periphery resets the pixel, which causes the pixel to memorize the new log intensity value. Images (a) and (b) are taken from [Delbruck, 2012], (c) is taken from [Liu and Delbruck, 2010], (d) and (e) are taken from [Delbruck, 2008].*

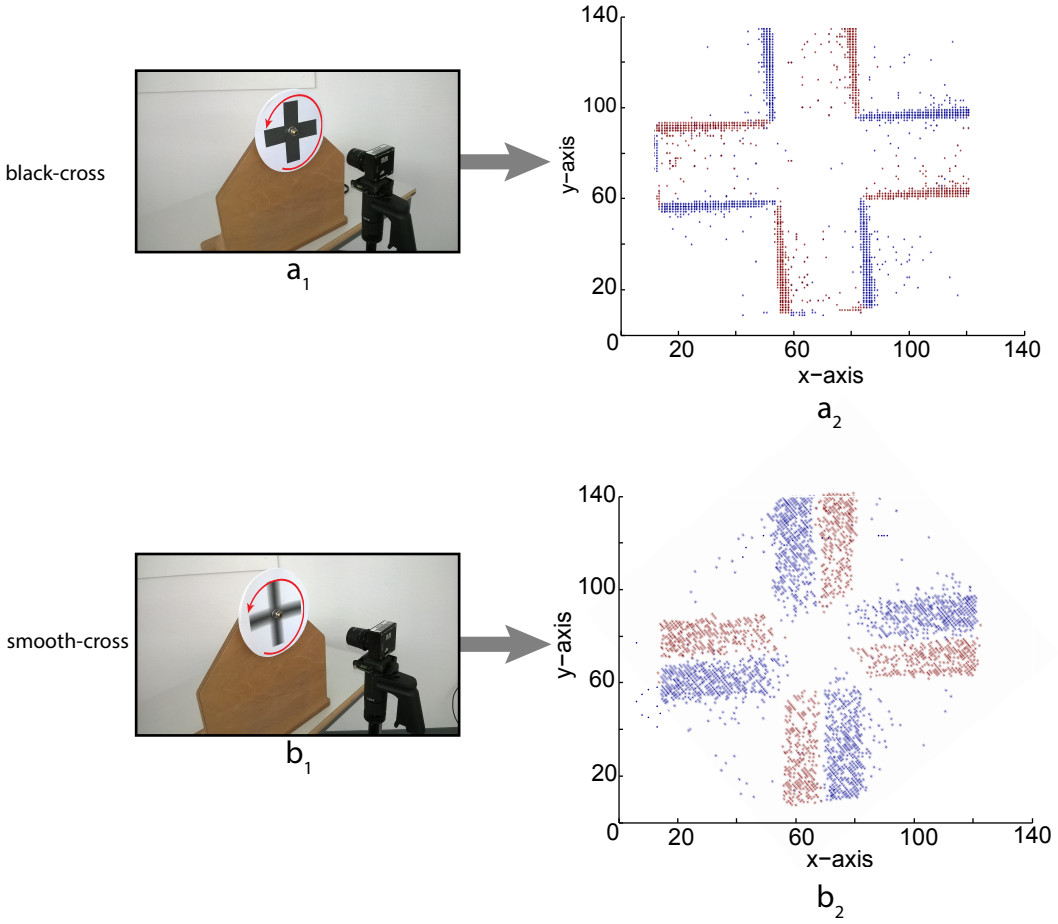


Figure 2.2: Examples of the DVS camera output for two stimuli, black-cross and smooth-cross. These stimuli were rotated counter-clockwise. (a_1) black-cross stimulus input, the DVS camera was mounted on a tripod and centered on the stimulus surface. (a_2) shows the ON (red points) and OFF (blue points) events that are generated via the DVS camera in response to the changes in intensity along the stimulus edges. (b_1) smooth-cross stimulus, again the DVS camera was mounted on a tripod and centered on the stimulus surface. (b_2) The output of the DVS camera where ON (red points) and OFF (blue points) events are generated over both edges and interior in response to the changes of the detected intensity over whole stimulus surface.

2.2.2 Input representation of ON/ OFF events

The DVS camera generates a stream of visual events that contains the spatial location (where the event occurs) and a time-stamp (when the event occurred). In order to provide a read out of this information, iniLabs has developed JAER software which is implemented in Java. This software enables an event from the DVS to be captured flexibly and presented on the monitor. Event information is provided as packets which can be defined as *Addr* and *T*. The first packet *Addr* includes information on the event address (the location in the spatial domain (x,y)) and event type if it is ON or OFF. This packet contains 15-bits in which the first 14-bits refer to the event address (each x and y address has 7-bits) which can be extracted by $x = 128 - V_{(1\rightarrow14)bits}$ and $y = 128 - (Addr_{(1\rightarrow14)bits} - x)/128$. The last bit in the packet *Addr* denotes the event label. This label is either one or zero, which in turn indicates an ON or OFF event, respectively. The second packet T directly specifies the time of the event with 1μ precision.

Figure 2.3 shows an output of the DVS128 camera in response to a rotated stimulus, namely a *half-circle stimulus*. The DVS camera was mounted on a tripod and centered on the stimulus surface. Figure 2.3(b) shows a table of actual data of T and *Addr* that were provided by the DVS camera in which the events position (x,y) and event types (ON/OFF) are extracted from the *Addr* packet. Since one event could not be visualized, we integrated the generated events in a 2-dimensional spatial domain (x,y), using different temporal windows as shown in Figures 2.3(c_1,c_2,c_3 and c_4). The number of integrated events increases with the increase in the temporal window. Figure 2.3(d) displays the ON and OFF events in the spatio-temporal domain(x,t). This figure shows how the DVS camera generates a stream of ON (reddish points) and OFF (bluish points) events that constitutes a kind of plait due to the rotational movement. To show the 3-dimensional construction of the ON/OFF events, we projected these events in a 3-dimensional space (x,y,t). This kind of visualization enables the 3-dimensional structure of a scene to be recovered and gives a general impression of how the event cloud in the 2-dimensional space (x,y) are depicted with respect to the t-axis.

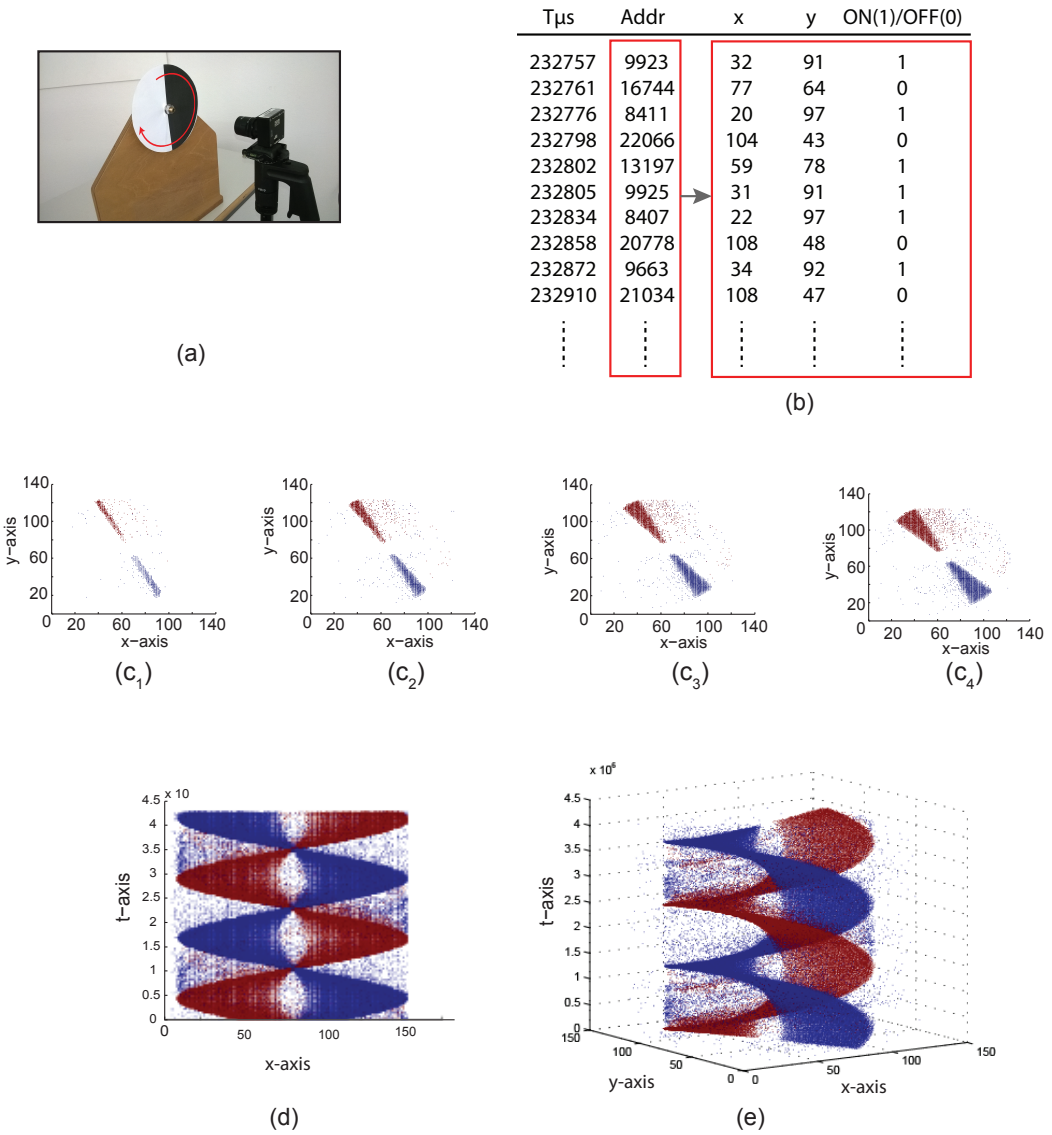


Figure 2.3: The output of the DVS128 camera. (a) The DVS camera placed on a tripod and centered on the rotated input image of half-circle stimulus using rotational actuator. The stimulus was rotated clockwise with speed of 5.23 rad/sec. (b) shows a table of actual data (T and Addr) that deliver form the DVS camera. The events position (x,y) and events types (ON/OFF) are extracted from Addr that are marked using a reddish square. (c_1) represents the ON (red points) and OFF (blue points) spikes responses that integrated within 20 msec while (c_2),(c_3) and(c_4) demonstrate the ON (red points) and OFF (blue points) events that integrated within larger temporal window 60 msec, 100 msec and 140 msec, respectively. (d) shows the generated stream of ON/OFF events in the spatio-temporal domain along (x,t) axes. (e) illustrates the events generated via DVS camera in 3-dimensional space (x,y,t).

2.3 Experimental devices and USB port limitations

Throughout our work, the synthetic ground truth data were generated using two rotational and translator devices, which are shown in Figure 2.4.

In the rotational recording, the desired stimulus was printed and arranged on a circular piece of paper with a diameter of 20 *cm*. The maximum input voltage of the rotational DC motors is 12 volts, which generates a maximum rotational speed of 6.28 rad/sec.

In the translator recording, the desired stimulus was printed and organized on a piece of paper of 20.5 $cm \times 73.3$ *cm*. This paper was rolled around two rotated cylinders, which fasten the paper and move it in a linear pathway. The maximum input voltage of the DC motors is 12 volts, which generates a maximum speed of 70 cm/sec.

In the rotational and translator motion recording, the DVS camera was mounted on a tripod in which the lens of the camera is directed to the center of the stimulus paper.

The emerging data, time-stamp and address-events, from the DVS sensor are transfered to a host PC over a USB port. Here, the interfacing USB port plays a critical rule in determining the amount of transmitted information between the visual sensor and the host computer. The DVS represents each event by 32-bits (4-bytes), which contains the time-stamp and address-event definitions. Since USB 1.0 is capable of 1.5 Megabytes per second and USB 1.1 can handle up to 12 Megabytes per second, the maximum amount of passing events is about 375×10^3 and 3×10^6, respectively. Since the DVS has a spatial resolution of 128×128, a fast stimulus that almost covers the visual field of the sensor could generate many activations exceeding the USB's capability [Lichtsteiner et al., 2008]; [Delbruck et al., 2010]. This problem can be overcome with USB 2.0 in which the available bandwidth of USB 2.0 is 480 Megabytes per second, which increases the capability of passing more events per second. We used a PC of Intel's (R) i5 - M460 processor (Core (TM) operating at 2.53 GHZ) and 8 GB memory (operating system: 64-bit windows 7).

2.4 Event-based vision sensor applications

High temporal resolution, low latency and large dynamic range visual sensing are key features of the address-event-representation (AER) principle, where each pixel of the vision sensor responds independently and almost instantaneously translates local intensity changes in the scene into events (ON or OFF).

A DVS quantizes local relative intensity changes to generate spike events with a low latency and large dynamic range. An event can be produced with a high temporal resolution (1 μs time precision), a wide dynamic range of illumination $120dB$, and a low latency of $15\mu s$ [Lichtsteiner et al., 2008].

The new technology of non-frame-based vision sensors has led to several recent applications in many fields in order to exploit the advantages of DVSs compared with traditional frame-based imagers, where several application-oriented studies have capitalized on those features. Such works include [Litzenberger et al., 2006a] and [Delbruck and Lichtsteiner, 2008], where Litzenberger and co-authors introduced an algorithm that used the silicon retina camera to estimate vehicle speed based on the slope of the events cloud. Delbruck and co-authors presented a hybrid neuromorphic procedural system for object tracking via an event driven cluster tracker algorithm in which a moving ball was considered as an example. The authors showed how the ball can be detected, tracked and successfully blocked by a goalie robot despite a low contrast object and complex background. The event-cluster algorithm was introduced by [Litzenberger et al., 2006b] and [Ni et al., 2011], whose first study considered a real world application, namely vehicle tracking for traffic monitoring in real time, and whose second study addressed microrobotics tracking.

In [Lee et al., 2012] and [Lee and Delbruck, 2014] Lee and co-authors proposed a real-time hand gesture interface based on a stereo pair of DVS silicon retinas. The authors detected motion trajectories of moving hands by spatiotemporally correlating the stereoscopically verged asynchronous events of DVSs using leaky integrate-and-fire (LIF) neurons. Lee and Delbruck presented a postprocessing framework based on spiking neural networks that can process the events received from the DVSs in real time, and which provides an architecture for future implementation in neuromorphic hardware devices. In [Drazen et al., 2011] the authors presented a proof-of-concept study examining the utility of using a DVS sensor for particle tracking measurements. Drazen and co-authors showed how the DVS vision sensor is able to seamlessly observe a moving small particle from one pixel to other with a limited bandwidth.

A description of an address-event vision system to detect accidental falls in elderly care homes was introduced in [Fu et al., 2008]. The system raises an alarm when a fall hazard is detected. [Delbrück and Lang, 2013] showed how an asynchronous neuromorphic dynamic vision sensor (DVS) silicon retina can be used to build a fast self-calibrating robotic goalie, which offers high update rates and low latency at low CPU load. In [Conradt et al., 2009] the authors demonstrated how a pair of spike-based silicon retinas (DVSs) were used to provide fast visual feedback for controlling an actuated table to balance an

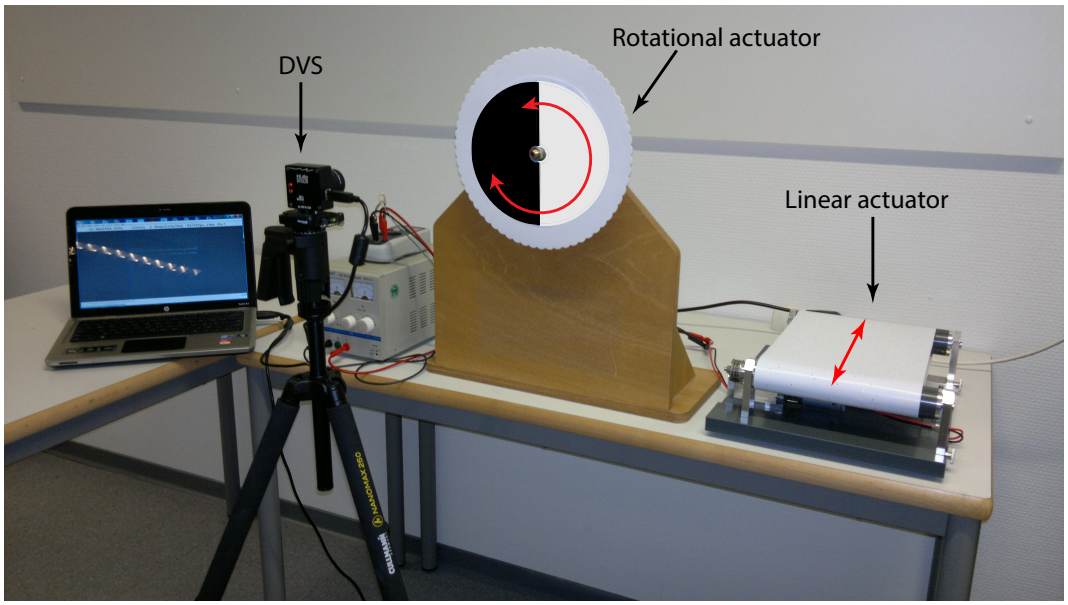


Figure 2.4: The experimental devices that used in our recording. The DVS is mounted on a tripod and the lens of the camera is directed to the center of the stimulus. The rotational actuator operates with 12 Volt as maximum input voltage that generates a maximum rotational speed 6.28 rad/s. The size of the diameter of the printed stimulus is 20 *cm. The maximum input voltage of the DC motors for the linear actuator is 12 Volt that drives a maximum speed 70 cm/sec. The desired stimulus was printed and organized on a a piece of paper size* 20.5 *cm* × 73.3 *cm. This paper is rolled and inserted in the devise that tighten the paper and move it with a linear pathway.*

ordinary pencil on its tip. Epipolar geometry has been studied by [Benosman et al., 2011]. The authors used stereo vision based events in which the geometric locus of the epipolar constraint can be measured directly without any prior knowledge of the binocular system. They showed how precise temporal and spatial locations are key features for a binocular system computational principle. Stereo vision based events for dynamic 3D visual reconstruction were investigated in [Camunas-Mesa et al., 2014]. The authors used a match technique for the corresponding events that were generated by two different sensors in a stereo setup. In [Gibson et al., 2014], artificial spiking neural networks (SNNs) were investigated by predicting future motion trajectories using an SNN with a deep learning architecture. The next subsection introduces the state-of-the-art methods for motion estimation based on event representation.

2.4.1 Motion estimation based on AER principles

Motion estimation is an advanced topic in automated visual processing and has been investigated widely using conventional cameras (see, e.g., [Strout et al.,]; [Lucas and Kanade, 1981]; [Brox et al., 2004] and [Drulea and Nedevschi, 2013]). Few studies have been published using the new vision sensor technology of an address-event silicon retina. In general, these studies fall into two categories: gradient-based and filter-based. The following subsections outline both categories along with the state-of-the-art literature.

2.4.1.1 Gradient based-event representation

Gradient techniques operate on the assumption of the *brightness change constraint* in which the intensity of the image remains constant within a small neighborhood in space and time. Such an assumption has been examined in [Horn and Schunck, 1981, Barron et al., 1994] (see Section 1.4.1 for discussions of the underlying assumption). Gradient techniques, also known as differential techniques, have been addressed by Benosman and co-authors [Benosman et al., 2012], who used the energy minimization method introduced in [Lucas and Kanade, 1981] to calculate the motion. Benosman and co-authors used the cloud of ON/OFF events generated by the silicon retina (DVS) as a visual input. As already mentioned, the main assumption of the motion estimation for such techniques is: the intensity function (I) captured by the retina sensor is invariant over a small surface patch. This assumption can be denoted by

$$\frac{d}{dt}I(x,y,t) = I_x u + I_y v + I_t, \tag{2.2}$$

where I_x,I_y and I_t represent the partial derivatives of the intensity function I with respect to x, y and time t, respectively. The parameters u and v define the local flow (velocity) vector $u = \frac{\partial x}{\partial t}$ and $v = \frac{\partial y}{\partial t}$. Intensity constancy assumption $\frac{d}{dt}I(x,y,t) = 0$ is more suitable with the functionality of AER sensor than the conventional frame-based sensor due to the low latency $15\mu s$ of the event-based sensors (see [Brosch et al., 2015]). Since the intensity constancy function Equation 2.2 contains two variables $(u,v)^T$ in one equation, an appropriate strategy needs to be used to overcome this problem. To tackle this problem, Benosman and co-authors use the suggestion in [Lucas and Kanade, 1981] where the velocity vector (u,v) is assumed to be constant over the neighborhood of the pixel (x,y). As a consequence, for $m \times m$ window, there are $n = m^2$ optical flow equations that can be written as

$$\mathbf{A} \cdot \mathbf{V} = \mathbf{b}, \tag{2.3}$$

$$\mathbf{A} = \begin{pmatrix} I_{x_1} & I_{y_1} \\ . & . \\ . & . \\ I_{x_m} & I_{y_m} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} u \\ v \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} -I_{t_1} \\ . \\ . \\ . \\ -I_{t_m} \end{pmatrix}$$

where I_x and I_y denote the spatial gradient over the x and y axes, respectively, while I_t represents the temporal derivative over the t axis. These equations can be solved using the least squares principle in which the solution of $\mathbf{V}$ is defined by $\mathbf{V} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$ (see [Lucas and Kanade, 1981]). Since the vision sensor generates a stream of events (ON or OFF) and does not provide gray levels, Benosman and co-authors suggested estimating the spatial and temporal gradient by comparing pixel activities over one temporal window to calculate the spatial gradient, and two temporal windows to calculate the temporal gradient using a short temporal window Δt, as shown below

$$\frac{\partial e(x,y,t)}{\partial x} \cong \sum_{t-\Delta t}^{t} e(x,y,t) - \sum_{t-\Delta t}^{t} e(x-1,y,t), \tag{2.4}$$

$$\frac{\partial e(x,y,t)}{\partial y} \cong \sum_{t-\Delta t}^{t} e(x,y,t) - \sum_{t-\Delta t}^{t} e(x,y-1,t) \tag{2.5}$$

$$\frac{\partial e(x,y,t)}{\partial t} \cong \frac{\sum_{t-\Delta t}^{t_1} e(x,y,t) - \sum_{t-\Delta t}^{t} e(x,y-1,t)}{t-t_1}, with t_1 < t. \tag{2.6}$$

Benosman and co-authors showed beneficial results, however, their method to approximate local gradients of the luminance function from event-sequences has its limitations and in some cases leads to inconclusive results [Tschechne et al., 2014a]. In a theoretical study of [Brosch et al., 2015] the authors showed that such an approach has multiple issues in terms of any real implementation. The most important observation was that when a luminance edge passes a pixel's receptive field of the DVS sensor, the number of events is in the range of about 10 (depending on the contrast, speed and luminance condition).

2.4.1.2 Filtering based-event representation

In the filtering technique, motion is estimated in the Fourier domain in which the optic flow of a scene is obtained based on spatiotemporal filters. The optic flow estimation based-filtering technique was recently presented by [Tschechne et al., 2014b], who used an events cloud as a visual input. In order to estimate a local motion flow calculated for each event occurring in the scene, Tschechne and co-authors implemented spatiotemporal filters of the type suggested by

the findings of [De Valoisa et al., 2000].

De Valois and co-authors showed that directionally-selective V1 simple cells have oriented RF profiles in the space-time domain, where in most cases these RFs can be constructed by a linear combination of a fast temporally-biphasic input and a slower monophasic input [DeValois and Cottaris, 1998]. These observations lead [Tschechne et al., 2014b] to the implementation of a family of spatio-temporally filters using the quadrature of spatial weighting profiles (odd and even) with mono and bi-phasic temporal filters. In their model, the spatiotemporal filters were implemented over a spatial buffer of (11×11) which stores the timestamp of the events.

Tschechne and co-authors extended their work in [Brosch et al., 2015] to investigate the implications of non-frame-based vision sensors in the context of visual motion. The authors discussed different principal approaches for motion estimation and showed that gradient-based methods for local motion estimation suffer from the sparse encoding in AER. On the other hand, [Brosch et al., 2015] showed that approaches exploiting the local plane-like structure of the events cloud are shown to be well suited.

The following sections introduce our model for motion estimation based on sparse events representation. Here, the spatiotemporal filters suggested by [Adelson and Bergen, 1985] are adaptive to be consistent with the AER principles.

2.5 Bio-inspired model for motion estimation using an address-event representation

In the previous section we demonstrated the recent findings for motion estimation based on sparse events representation. Such findings showed that the sparse events distribution can be appropriately defined through the local filtering structure such as the spatio-temporal RFs of V1 cells. In this work, we propose a neural mechanism for motion estimation based on the filtering structure. Our model uses the events cloud distribution generated by the silicon retina, DVS. This sensor requires an adaptive model for motion estimation that is consistent with the functionality of the AER. As mentioned in Section 2.2.1, the silicon retina generates an asynchronous stream of spike events based on quantizing local relative intensity changes. The DVS thus operates based on temporal derivatives of the input signal. This feature was considered in our model where we modify the functionality of Adelson-Bergen spatio-temporal filtering to create a representation of the cloud events input.

In our model, we propose a neural mechanism for initial motion estima-

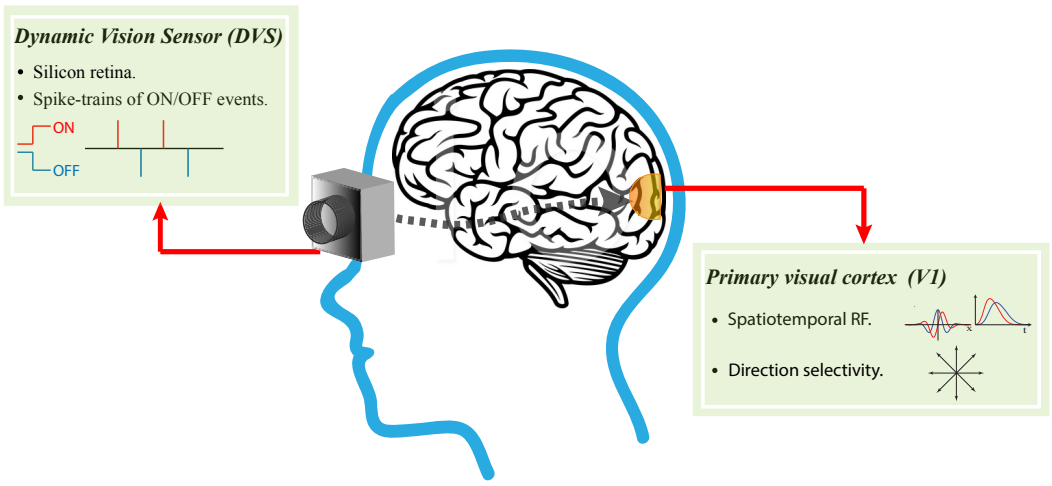


Figure 2.5: Overview of the proposed model structure. The visual scene is captured using the silicon retina (DVS) in which spike-trains of ON/OFF events are generated. The local motions of the sparse of input events are estimated via the primary visual cortex (area V1) where the visual input data are analyzed. The RFs profiles of cells in area V1 are modeled using spatio-temporal filters.

tion which is functionally related to the dorsal stream in the visual cortical hierarchy based on sparse input representation. We adapt the spatio-temporal filtering scheme as originally suggested by Adelson and Bergen to make it consistent with the input representation generated by the DVS. Our model extracts motion information via a spatiotemporal energy mechanism, which is oriented in the space-time domain and tuned in spatial frequency. In addition, our model introduces subsequent competitive interaction among the responses of the filtering models to ensure that the activities of individual cells are balanced against the neighborhood activities.

The rest of this chapter focuses on the computational mechanisms used in our model, which is biologically inspired. Figure 2.5 shows the visual processing areas adopted along the dorsal pathway. Here, the DVS sensor emulates the main processing sequence of the mammalian retina by generating spike-trains to encode temporal changes in the luminance distribution of a visual scene. These spike-trains are transferred to the primary visual cortex (V1), in which spatio-temporal filtering mechanisms are used. The following subsections detail the methodology of our model.

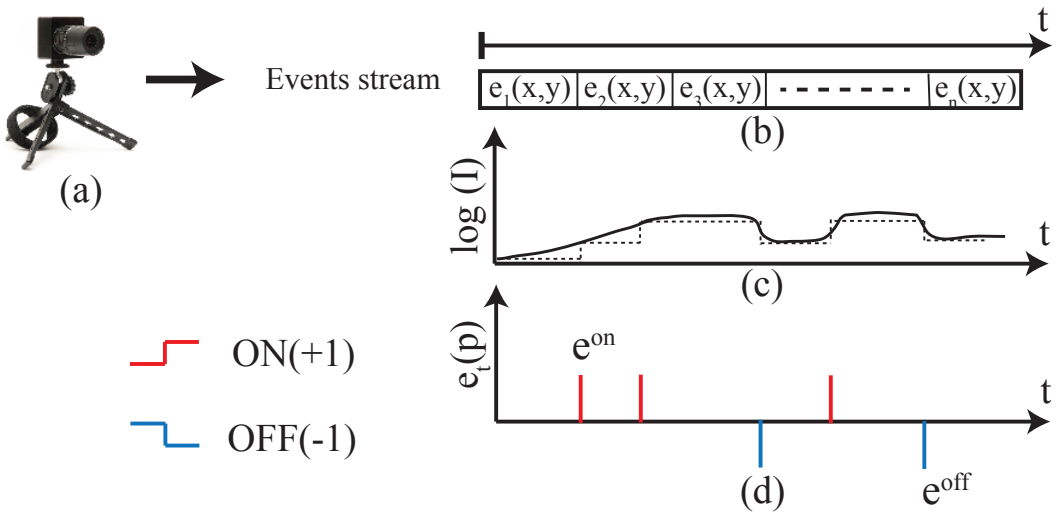


Figure 2.6: Initial input stage. (a) DVS sensor. (b) Event stream which is represented as a sequence of events e at a position (x, y) and time t. (c) Local changes in intensity (log I) excite (ON or OFF) events. (d) e^{on} and e^{off} identify the event activity (+1) ON and (-1) OFF, respectively.

2.5.1 Initial input stage (silicon retina)

In the initial input stage, DVS sensor is used to capture the perceiving visual scene. As mentioned in Section 2.2.1, the neuromorphic vision sensor generates spikes or events, where each pixel of the vision sensor responds independently and almost instantaneously translates local changes in the intensity (log I) into events (ON or OFF), see Figure 2.6. This principle is used in our study to profit from the advantage of the event-based technology instead of using standard frame-based camera technology. Since a single event in spatiotemporal domain yields an ambiguity for motion estimation using spatiotemporal filtering, pixel activities, ON (+1) and OFF (-1), are accumulated during a temporal window, Equation 2.7. The interval length of the temporal window can be parametrized in accordance with the extent of the temporal impulse responses of the filters. Activity accumulation of ON and OFF events is thus given as

$$e(x, y, t) = e^{on}(x, y, t) + e^{off}(x, y, t), \tag{2.7}$$

where $e^{on}(x, y, t)$ and $e^{off}(x, y, t)$ represent ON and OFF events respectively, occurring at position (x, y) and time t.

2.5.2 Motion detection stage (area V1)

In this section, we introduce how the spatiotemporal filters that were suggested by [Adelson and Bergen, 1985] are adapted to be compatible with the DVS sensor. In principle, the spatiotemporal filters can be constructed by the combination of two separable filters with a specific spatiotemporally characteristics. This has been investigated by physiological scientists (see [Watson and Ahumada, 1983]) where [Adelson and Bergen, 1985] introduced a wide variety of filter combinations to generate spatiotemporal oriented responses. The main idea of the spatiotemporal orientation is summing the responses of four separable filters, which are identical except for a shift in receptive-field center and a temporal delay.

Motion estimation using spatiotemporal filters emulates motion detection processing of the primary visual cortex [Adelson and Bergen, 1985], where the space-time filters are 3D and can here be decomposed into separable products of two 2D spatial and two 1D temporal kernels. The temporal filters use two different integration windows (fast and slow), while the spatial filter uses two different phases (even and odd). These filters are modified to be consistent with the spike of events representation. The following subsections define the structure of the employed spatiotemporal filters in our study.

2.5.2.1 Spatial filters

The spatial receptive fields (RFs) of odd and even filters can be implemented using Gabor functions, which provide a close description of the receptive fields in primary visual cortex area (V1) [Ringach, 2002]. We thus used these functions to build even and odd spatial filters as in Eqs.(2.8) and (2.9), respectively namely

$$F_{even}(x, y, \theta_k, f_s) = \frac{1}{2\pi\sigma_s^2} \cdot exp(-\frac{\breve{x}^2 + \breve{y}^2}{2\sigma_s^2}) \cdot cos(2\pi f_s \breve{x}), \tag{2.8}$$

$$F_{odd}(x, y, \theta_k, f_s) = \frac{1}{2\pi\sigma_s^2} \cdot exp(-\frac{\breve{x}^2 + \breve{y}^2}{2\sigma_s^2}) \cdot sin(2\pi f_s \breve{x}), \tag{2.9}$$

where $\breve{x} = xcos\theta_k - ysin\theta_k$, $\breve{y} = xsin\theta_k + ycos\theta_k$, θ_k is the spatial filter orientation with N different orientations where $k = \{1, 2, 3...N\}$, σ_s is the standard deviation of the spatial filters and f_s represents the spatial frequency tuning.

2.5.2.2 Temporal filters

In [Adelson and Bergen, 1985], the authors used temporal gamma functions ($f(t) = (kt)^2 \cdot exp(-kt^2) \cdot [1/n! - (kt)^2/(n+2)!]$) of different durations (n=3, n=5) in order to achieve temporal smoothing and differentiation. Since event-based sensor responses already encode temporal luminance changes, i.e. temporal derivatives of the input signals, we employ a convolution process utilizing smoothing temporal filters (Equation (2.10)) which integrate the input stream of events.

$$f(t) = \int (kt)^2 \cdot exp(-kt^2) \cdot [1/n! - (kt)^2/(n+2)!]. \tag{2.10}$$

This allows us to obtain scaled versions of temporally smoothed derivatives of the input luminance function as input representation for motion estimation. To simplify the mathematical description of the temporal filters in Equation (2.10), we suggest to combine two Gaussian functions (Λ) having different standard deviations(σ) and mean values (μ). This combination is scaled by a scale factor c to closely resemble the shape of the Adelson-Bergen temporal filters as given in Equation (2.11)

$$S_{slow,fast}(t) = \left[\Lambda_{\sigma_1,\mu_1}(t) - \Lambda_{\sigma_2,\mu_2}(t)\right]/c, \tag{2.11}$$

with $\Lambda_{\sigma_*,\mu_*}(t) = 1/(\sigma_*\sqrt{2\pi}) \cdot exp(-(t-\mu_*)^2/(2\sigma_*^2))$, where μ_* and σ_* represent the mean and the width of temporal window extent of the Gaussian function, respectively. Figure 2.7 shows the temporal filters (fast and slow) of Adelson-Bergen and the approximate profiles of the Gaussian functions combination. We used $\sigma_1 = 1$, $\mu_1 = 2.5$, $\sigma_2 = 2$, $\mu_2 = 7$, $c = 2.6$ to generate the fast temporal filter and $\sigma_1 = 1.3$, $\mu_1 = 4$, $\sigma_2 = 2.3$, $\mu_2 = 9.2$, $c = 3.1$ to generate the slow temporal filter. Subsequently, the smoothing temporal filters $T_{slow,fast}(t)$ are calculated based on the integral $\int S_{slow,fast}(t)dt$, Equation (2.11), and then the summation of each filter is scaled to 1 to prevent any biases in calculating responses. The smoothing temporal filters are thus given by

$$T_{slow,fast}(t) = \frac{1}{\int_0^\infty (G_1 - G_2)dt} \cdot (G_1 - G_2)(t), \tag{2.12}$$

$$G_1 = \left[\frac{1}{2}(1 + erf\left(\frac{(t-\mu_1)}{\sigma_1\sqrt{2}}\right)\right]/c, \tag{2.13}$$

$$G_2 = \left[\frac{1}{2}(1 + erf\left(\frac{(t-\mu_2)}{\sigma_2\sqrt{2}}\right)\right]/c, \tag{2.14}$$

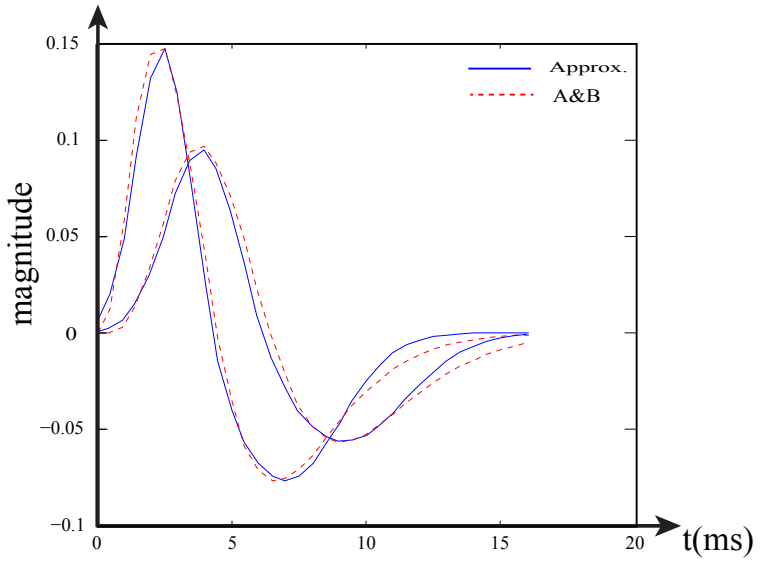


Figure 2.7: Fast and slow temporal filters. Dashed lines show Adelson-Bergen (A&B) filters and solid lines show the approximated shapes (Approx.) derived by two shifted Gaussian envelopes as filters.

where $T_{slow,fast}(t)$ defines the fast and slow temporal filters, G_1 and G_2 represent the integral of Gaussian functions $\int \Lambda_{\sigma_1,\mu_1}(t)dt$ and $\int \Lambda_{\sigma_2,\mu_2}(t)dt$, respectively.

2.5.2.3 Combination of spatial and temporal filters

In this part, we show how the spatial and temporal filters can be combined to generate directionally selective filters. Figure 2.8 (a) and (b) show the spatial filters ($F_{even}(x, y, \theta_k, f_s)$, $F_{odd}(x, y, \theta_k, f_s)$) and the temporal filters ($T_{slow,fast}(t)$), respectively. The spatiotemporal separable filters were calculated according to the scheme proposed in [Adelson and Bergen, 1985] in which the products of two spatial and two temporal filters are shown in the first row of Figure 2.8 (c). These filters are combined in a linear fashion in order to obtain the oriented selectivity in the spatio-temporal domain, as shown in the second row of Figure 2.8 (c). The oriented linear combinations are denoted by

$$\begin{aligned} F_a^{v1}(x, y, \theta_k, f_s, t) &= F_{even}(x, y, \theta_k, f_s) \cdot T_{slow}(t) + \\ & \quad F_{odd}(x, y, \theta_k, f_s) \cdot T_{fast}(t), \end{aligned} \tag{2.15}$$

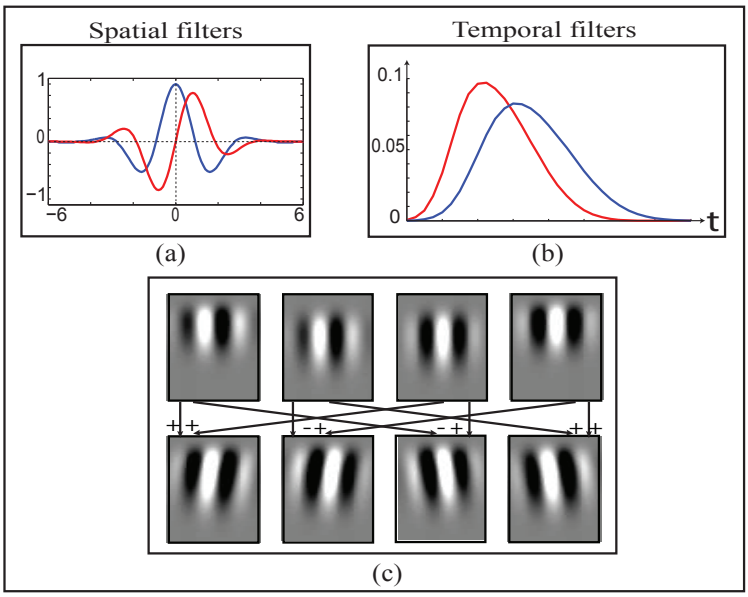


Figure 2.8: Spatiotemporal filter construction. (a) Spatial filters. (b) Temporal filters. (c) The first row represents the products of two spatial and two temporal filters. The second row represents the sum and difference of the spatio-temporal product filters.

$$F_b^{v1}(x, y, \theta_k, f_s, t) = F_{even}(x, y, \theta_k, f_s) \cdot T_{fast}(t) - F_{odd}(x, y, \theta_k, f_s) \cdot T_{slow}(t). \quad (2.16)$$

The filtering response of area V1 for a stream of events input $e(x, y, t)$ can be calculated by the gradual activation response function $\dot{r}^{v1} = -r^{v1} + ([F_a^{v1} * e]^2 + [F_b^{v1} * e]^2)$ which has been solved at equilibrium state

$$r_\theta^{v1} = ([F_a^{v1} * e]^2 + [F_b^{v1} * e]^2). \quad (2.17)$$

The symbol $'*'$ indicates the convolution operator, θ indicates motion directions (left vs. right relative to the orientation axis). For better readability we omitted the local spatial coordinates and feature selectivities.

Figure 2.9 reviews the stages of our suggested model. In this figure, we followed the structure of Adelson and Bergen (see Figure 1.7) in which the main stages that simulate the responses of area V1 have been demonstrated. In the first stage, the ON/OFF events stream $e(x, y, t)$ is fed to the second and third stages where two spatial filters (F_{even}, F_{odd}) and two temporal filters (T_{slow}, T_{fast}) are utilized. As a result, four separable combination filters are generated A, A', B, B' in the third stage. Sums and differences of the four separable responses are taken in the fourth stage to produce the spatiotemporally oriented linear responses. These responses are selective for direction of motion (rightward and leftward) that define the responses of the V1 simple cells. In the fifth stage, response pairs for rightward and for leftward motion are combined by summing their squares, i.e., a nonlinear combination of V1 simple cells. In this stage the oriented-energy responses are derived which define the characteristics of V1 complex cells responses.

2.5.3 Response normalization

The activity in neurons show significant nonlinearities depending on spatiotemporal activity distribution in the space-feature domain surrounding a target cell [Carandini and Heeger, 2012]. Such response nonlinearities have been demonstrated in the LGN, early visual cortex (area V1), and beyond. In theoretical studies (e.g. [Brosch and Neumann, 2014]) it has been proposed that such compression of stimulus responses can be achieved through the normalization of the target cell response defined by the weighted integration of activities in a neighborhood defined over the space-feature domain of a target cell. In other words, the normalization operation utilizes contextual information from a local neighborhood that is defined in space as well as feature domains relevant for the current computation. Such a normalization can be generated at the neuronal level by divisive, or shunting, inhibition (see [Blomfield, 1974], [Dayan

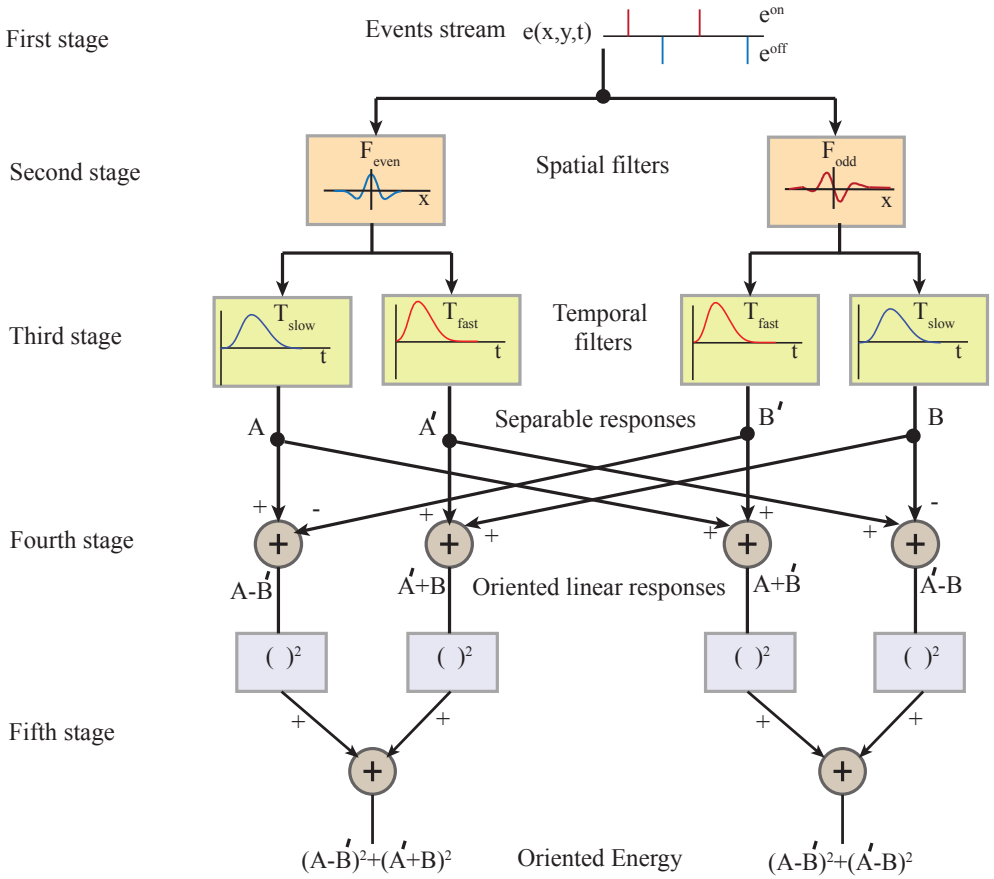


Figure 2.9: The stages of our suggested model following to the Adelson-Bergen energy model. First stage, shows the input of ON/OFF events stream $e(x, y, t)$ at position (x, y) and time t. Second stage, the events stream passes through the two spatial filters with different phases (even and odd) F_{even} and F_{odd}. Third stage, each output is fed to two temporal filters with different integration window T_{slow} and T_{fast} lead to drive four separable responses A, A', B, and B'. Fourth stage, the four separable responses are summed and subtracted to generate directionally selective filters in which oriented linear responses are generated. Fifth stage, squares of quadrature pairs of the oriented linear responses are summed to generate motion energy for each direction.

and Abbot, 2001], [Silver, 2010].

Given the activity of a neuron (defined by its membrane potential) the rate of change can be characterized by the following rate equation [Grossberg, 1988]

$$\tau \frac{dv\,(t)}{dt} = -A \cdot v\,(t) + (B - C \cdot v\,(t)) \cdot net_{ex} - (D + E \cdot v\,(t)) \cdot net_{in}, \quad (2.18)$$

given A representing the passive leakage, B and D are parameters to denote the saturation potentials (relative to C and E, respectively), and net_{ex} and net_{in} denote generic excitatory and inhibitory inputs to the target cell and impose the conductances of the membrane potential. In order to achieve balanced cell activations against the pool of neighboring cells, a normalization is calculated, following [Bouecke et al., 2011] and [Brosch and Neumann, 2014]. We employ a spatial weighting function Λ_σ^{pool} which realizes a distance-dependent weighting characteristics, e.g., Gaussian. The size of this neighborhood function is larger than the receptive field, or kernel, size of the cells under consideration. After normalization of activations, the responses are guaranteed to be bounded within a local activity range. In addition, a spectral whitening of the local response distribution occurs [Lyu and Simoncelli, 2009].

We realized a slightly simplified version of the scheme described in [Brosch and Neumann, 2014] and solve the normalization interaction at equilibrium, namely evaluating the steady-state response for $\frac{dv(t)}{dt} = 0$. Even further, we set C = 1 and D = 0 in Equation (2.18) to define shunting inhibition, and B = E = 1 to scale the response levels accordingly. As a consequence, we get the steady state response for Equation (2.18) which reads

$$v_\infty = \frac{net_{ex}}{A + net_{ex} + net_{in}}. \quad (2.19)$$

We normalize the model responses r^{v1} in the spatial domain using an integration field that weights the activities in the spatial neighborhood of the target. We propose a spatial weight fall-off in accordance to a Gaussian weighting function. The motion selective responses are defined in direction space relative to the local contrast orientation θ of the spatial filter kernels used. We take the direction feature space into account as well by calculating the average activity over all directions. In all, we can denote the overall pool activation by

$$r^{pool}(x, y) = \frac{1}{2N} \sum_\theta \left\{ r_\theta^{v1} * \Lambda_\sigma \right\}_{\mathbf{x,y}}, \quad (2.20)$$

with θ denoting motion direction, $'*'$ denotes the (spatial) convolution operator, (x, y) represents the spatial position of the cell, N is the number of contrast filter orientations and Λ is the weighting function of the spatial pooling operation. The latter is parametrized by the parameter σ to denote the width of the spatial extent. The resulting normalized response is finally calculated by

$$r_\theta^{v1nor}(x,y) = \frac{r_\theta^{v1}(x,y)}{A + r_\theta^{v1}(x,y) + r^{pool}(x,y)}, \tag{2.21}$$

A denotes the passive decay of the dynamic mechanism and prevents from zero division in the steady-state.

2.6 Experimental setup and results

2.6.1 Ground truth data

To evaluate our method, a set of different stimuli with translatory and rotational motions were recorded using the DVS128 sensor. The rotational and translatory motions were generated using linear and rotational actuators, in which the linear actuator's speed is 20 cm/sec and the rotational actuator's speed is 5.23 rad/sec. The DVS sensor was mounted on a tripod and placed 23 cm away from the stimulus.

In our model, we used eight directions of motion (N =8 in which $\theta = 0°$, 45°, 90°, 135°, 180°, 225°, 270°, 315°). Here, the spatial filter frequency f_s is set to 0.25 and the standard deviation of spatial filters σ_s is set to 2 $pixel$. In the normalization mechanism, the standard deviation of Gaussian function σ is set to 15 $pixel$, while the leakage activities A is set to 0.01. The estimated results of the optic flow were based on the summed responses of r which generates a confidence for the motion direction $(u_e(x,y)\, v_e(x,y))^T = \sum_\theta r(x,y,\theta,f_s,t) \cdot (cos\,\theta \quad -sin\,\theta)^T$. Figures 2.10 and 2.11 present the translatory and rotational motion results respectively, where the stimulus image, ON/OFF events and ground truth are presented in the first row. The second row shows the estimated flow and the direction selectivity which is depicted in polar plot. In order to measure the accuracy of our approach, we calculated the angular error $\Phi(x,y) = cos^{-1}(\mathbf{V}_e(x,y) \cdot \mathbf{V}_g(x,y))/(|\mathbf{V}_e(x,y)||\mathbf{V}_g(x,y)|)$, where $\mathbf{V}_e(x,y)^T = (u_e(x,y), v_e(x,y))$ and $\mathbf{V}_g(x,y)^T = (u_g(x,y), v_g(x,y))$ represent the estimated and ground truth flow vectors at position (x,y), respectively. The error values in the range of $[0°, 180°]$ are depicted as a histogram which is shown in the third column.

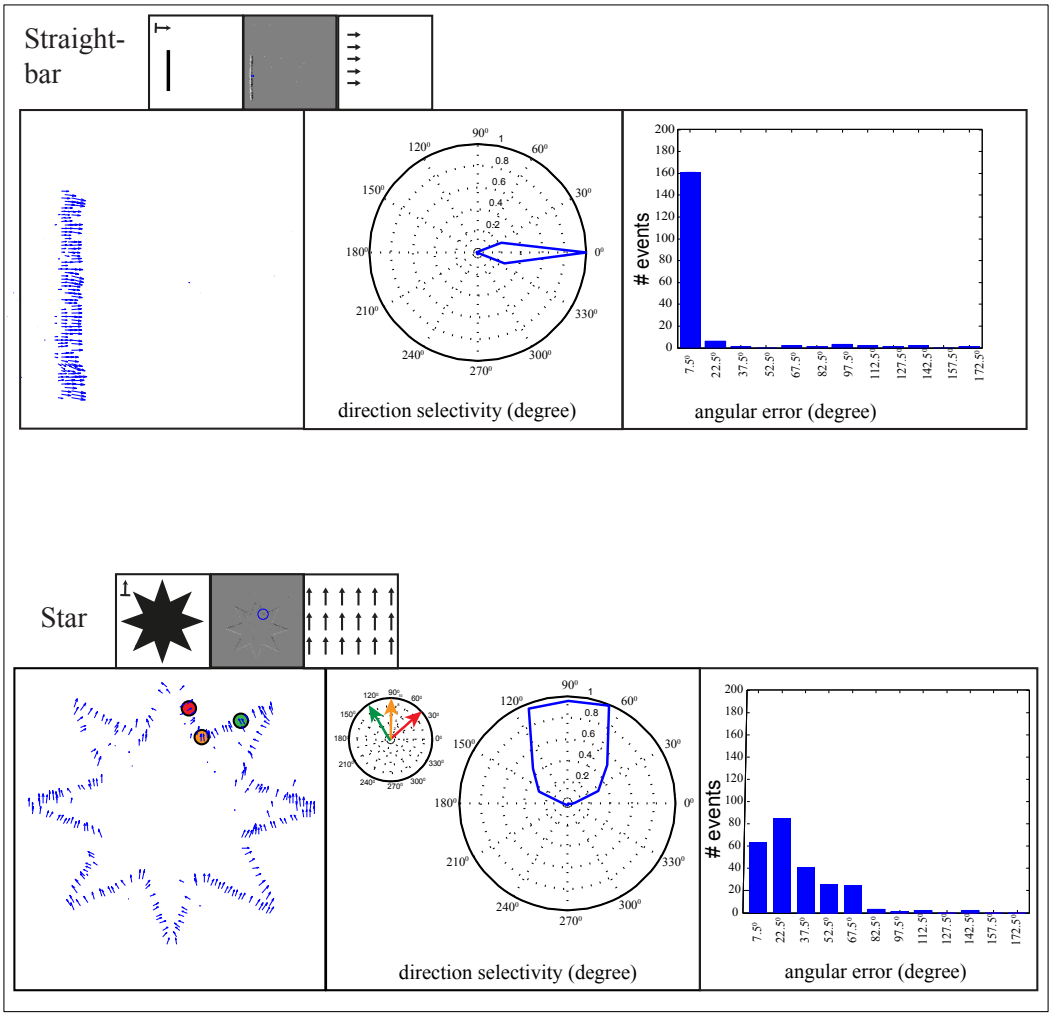


Figure 2.10: Processing results of translatory motion stimuli, straight-bar and star. The first row of each stimulus contains the input image, the ON/OFF accumulated events (3 msec.) and a sketch of the ground truth optical flow field. The second row shows the estimated motion, the polar plot of the direction selectivity and the histogram of the angular error between the estimated motion and their respective ground truth. In the star stimulus, the smaller polar plot represents the direction selectivity at corner and bar contours while the larger polar plot represents the direction selectivity of the whole stimulus. The abscissa of the histogram represents the binning in the range of the angular errors Φ that are combined into single bins $[\theta - 7.5°, \theta + 7.5°)$. The ordinate represents the number of pixels.

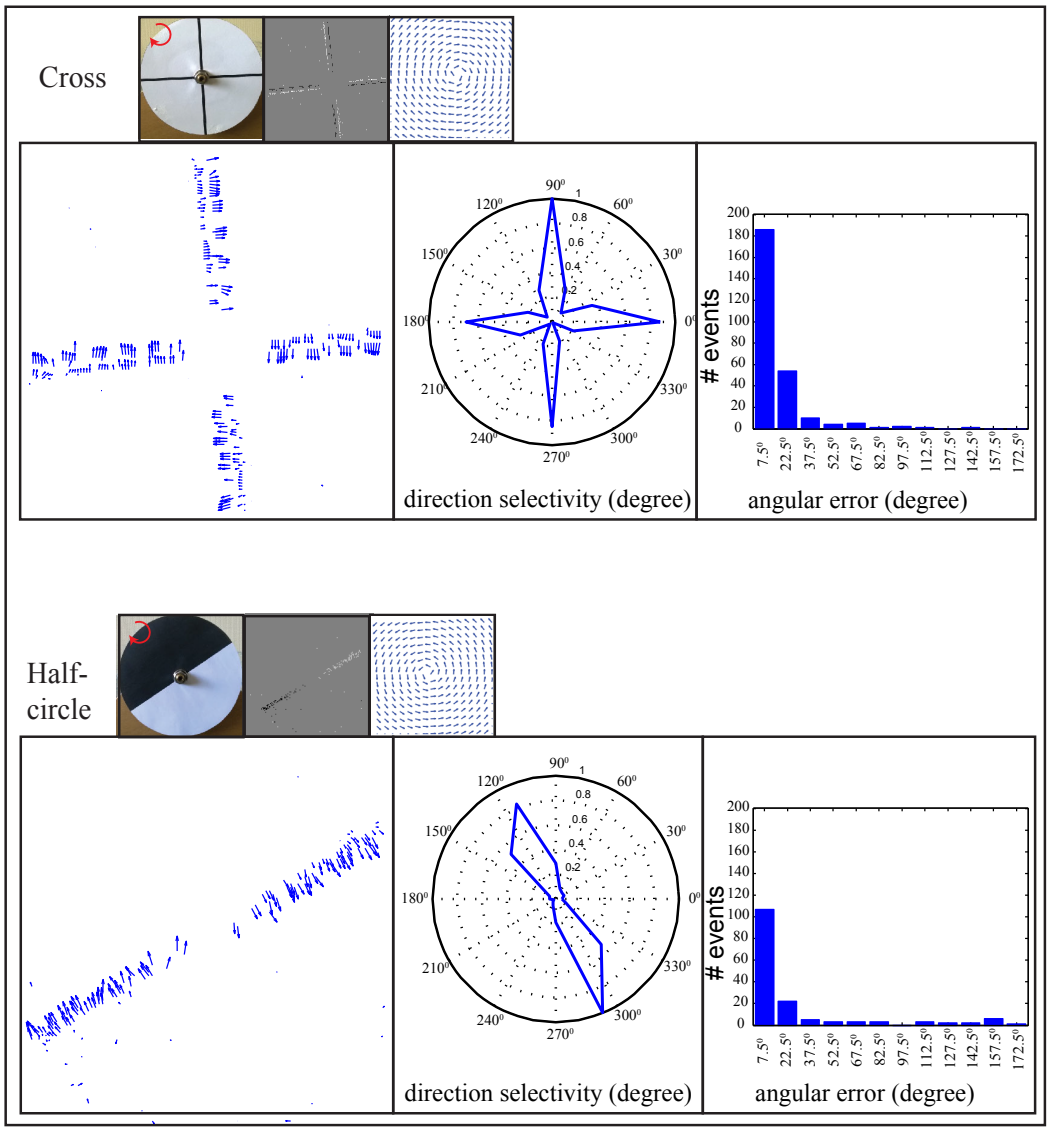


Figure 2.11: Processing results of clockwise rotational motion stimuli, cross and half-circle. The first row of each stimulus contains the input image, the ON/OFF accumulated events (3 msec.) and the ground truth optical flow field. The second row represents the estimated motion, the polar plot of the direction selectivity and the histogram of the angular error between the estimated motion and their respective ground truth. The abscissa of the histogram represents the binning in the range of the angular errors Φ that are combined into single bins $[\theta - 7.5°, \theta + 7.5°)$, The ordinate represents the number of pixels.

In case of *translatory* motion, we used two stimuli: a straight bar and a star in which different directions were selected to move these stimuli. The straight bar stimulus was moved to the left while the star was moved vertically up the field of view. For the straight bar, the polar plot shows that the spatiotemporal filters are selective to the right direction (0°). According to the histogram, the angular error between the estimated flow and the ground truth flow reveals that motions were estimated with correct directions in which the error values are accumulated within a small range of $[0°, 15°)$. In the star stimulus, different slanted bars are connected to form the star shape. Since each bar generates motion components that suffer locally from the aperture problem the initial estimation and subsequent integration impose a challenge. The smaller polar plot shows the direction selectivity at a particular position (corner and bar contours) while the larger polar plot shows the whole direction selectivity. Although an ambiguous motion was estimated along bar contours, real motion was estimated at the corner regions (aperture problem, see Section 2.7). The aperture problem can be resolved via the feedback of larger integration receptive field MT cells (for more details see [Bayerl and Neumann, 2004]). The angular error histogram showed some spurious flow in 45° due to the small slanted lines. Moreover, a smaller spurious flow occurred $\Phi > 60°$ due to the low resolution of DVS (128×128) which gives rise to spatial aliasing.

In the case of *rotational* motion, again two stimuli were used: cross and half-circle. These stimuli were rotated clockwise. In the cross stimulus, the motion estimation showed a flow pattern on the stimulus blades in which motion was estimated with preferred direction of (0°, 90°, 180°, 270°). In the half-circle stimulus, the results showed a flow motion over the stimulus diagonal along the elongated contrast edge contour. Here, the direction selectivity of the spatiotemporal filters is tuned to 290° and 110°. In both stimuli, the results revealed appropriate flow estimation compared with the ground truth of the clockwise rotation in which most of the angular error was sandwiched between 15° and 30°.

2.6.2 Complex realistic movements

To demonstrate the usefulness of our model under realistic acquisition conditions, we extended our evaluation to include bouncing ball and articulated biological motions, Figures 2.12 and 2.13. In the case of the bouncing ball, different projected velocities occur since the ball moves from a distant position towards the camera. On the other hand, the biological motions generate different speeds and directions inducing by the body and limbs movements. This leads to the additional challenge for our model to estimate the motion when

different velocities occur.

In the bouncing ball, we tested the influence of the temporal window size of events integration on the motion estimation. We used three different window sizes, $20msec$, $10msec$ and $3msec$, in which the first accumulated temporal period is equivalent to the typical sampling rate of a conventional frame-based imager. Figure 2.12 shows that the motion estimation can be improved with decreasing the temporal window size. This referred to the higher sampling rate interval can capture small number of events and instantaneously transcribe their motion in contrast to the larger interval window that integrates more events over time space which leads to lose the intermediate motion details. In general, the result of smaller number of events acquisition, Figure 2.12 (c), shows a proper estimation of flow direction comparing with other sampling cases.

Our model was tested using articulated movements in which real body motions are represented. Figure 2.13 shows two actions (jumping-jack and two hands waving) of an actor, where different movements and speeds were generated from body and limbs motions. The estimated motions for the two actions have been determined using a sampling rate of $3\,msec$ in which a satisfactory motions are obtained. As a further step towards the implication of the biological motion estimation, Chapter 4 demonstrates how representative features can be extracted from these motions which are then used for a classification purpose (biological motion recognition).

2.7 Aperture problem

Neurons in the primary visual cortical area V1 that are selective to spatio-temporal stimulus features have small RFs, or filter sizes. Consequently, they can only detect local motion components that occur within their RFs as shown in Figure 2.14 (a). This means that along elongated contrasts only ambiguous motion information can be detected locally. It is the normal flow component that can be measured along the local contrast gradient of the luminance function (aperture problem). Figure 2.14 (b) illustrates some of possible velocities generating at the center of a bar contour where these velocities are represented in the velocity space as shown in Figure 2.14 (c). In our test scenarios aperture problem has been investigated with input shown in Figure 2.8. The aperture problem can be resolved either by utilizing local feature responses at corners, line ends, or junctions that belong to a single surface to be tracked. Another approach is to integrate several normal flow estimates at distant locations. The following subsections review some of the common solutions of the aperture problem.

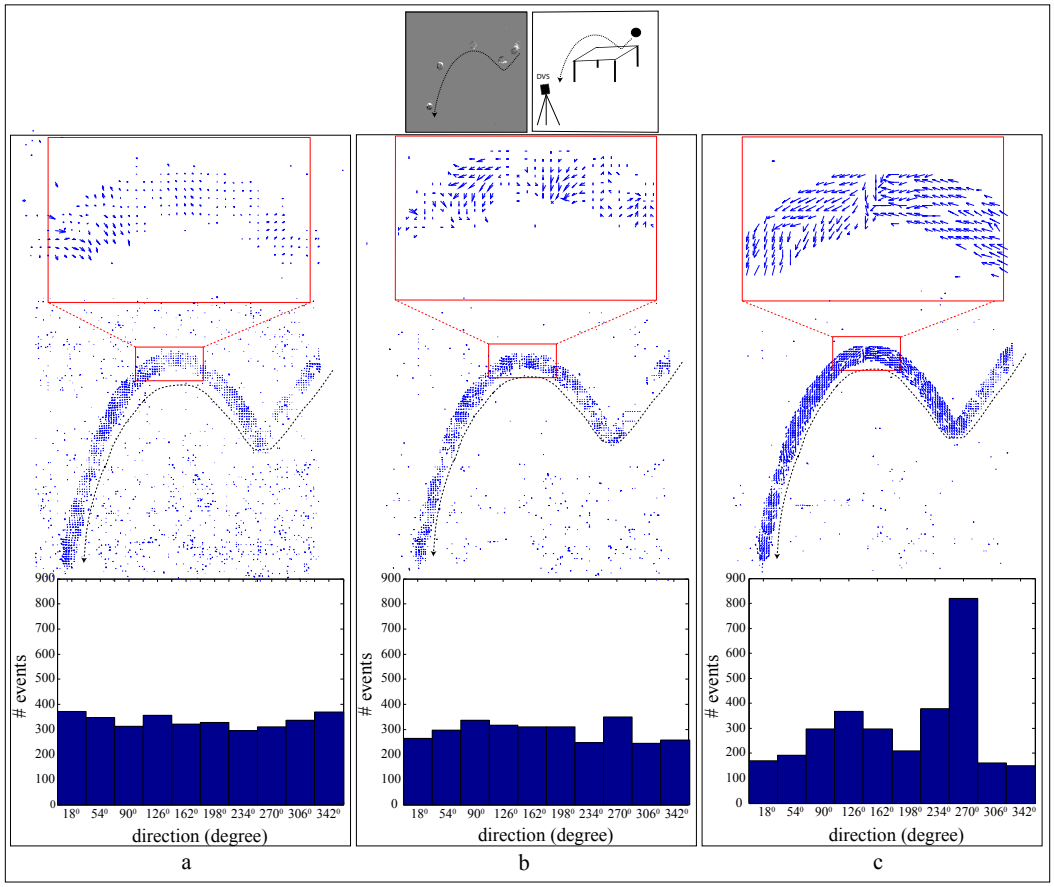


Figure 2.12: Estimated motion of bouncing ball: The first row depicts the experimental setup and samples of ball events. The second row shows the optic flow for the ball path. (a) Motion estimation using temporal window of $20\,msec$. *(b) Motion estimation using temporal window of* $10\,msec$. *(c) Motion estimation using temporal window of* $3\,msec$. *The second row shows the histograms of the estimated directions where the abscissa represents the binning in the rang of the direction* θ *that are combined into one bin* $[\theta - 18°, \theta + 18°)$, *and the ordinate represents the number of the events.*

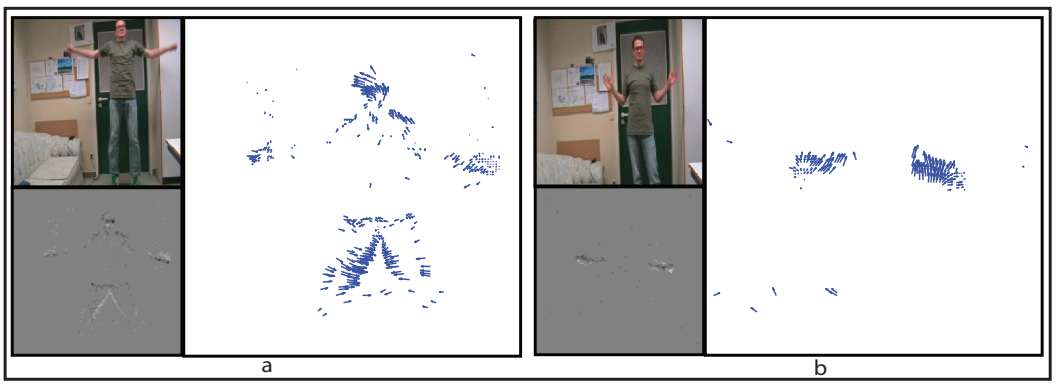


Figure 2.13: Estimated flow of articulated motion of an actor. (a) Jumping-jack. (b) Two hands waving. The first column of each action represents original image input and the integrated ON/OFF events. The second column of each action represent the estimated motions.

2.7.1 Intersection of Constraints and Vector Average

The intersection-of-constraints (IOC) [Adelson and Movshon, 1982] and vector average (VA) [Yo and Wilson, 1992] approaches can be demonstrated to calculate the exact movement of, e.g., two overlapping gratings which move translatory behind a circular aperture in distinct directions (plaid I and plaid II), see Figure 2.14 (d) and (e). If the two gratings have same contrast and spatial frequency the plaid appears as a single pattern. In the IOC mechanism the motion pattern of the two gratings is predicted in the direction of the intersecting normal flow constraint lines defined by the component gratings. This direction correspond to the feature motions generated by the grating intersections. While in the VA approach the motion pattern is predicted by averaging the motions of both gratings. In plaid I, the VA and IOC approaches produce motion pattern in the same direction since the moving directions of both gratings lie on either side with respect to the generated movement of 2D crossings. While in plaid II, gratings move in similar direction (i.e. both movements lie on the same side with respect to the generated movement of 2D crossings) that lead to different motion directions predicted by the VA and IOC approaches .

The IOC and AV methods could be used here as well by utilizing a voting scheme that is initialized by the normal flow components derived from the spatio-temporal filter responses as described above (in Section 2.5.2). Since the spatio-temporal weights of the filters already take into account the uncertainty of the detection and estimation process the IOC approach could be formulated within the Bayesian framework [Simoncelli, 1999]. To implement this

mechanism in our model, we could combine the local estimated motions from spatio-temporal filters and the likelihoods for the corresponding constraint lines. The IOC solution would then be the maximum likelihood response of the multiplied constraint component likelihoods. The integration of normal flow motions in the IOC (as well as the VA) is valid under the assumption that the contributions from component flows are generated by translatory motions. For rotational flows of an extended object, such as the ones shown in Figure 2.11, these approaches do not yield the correct integrated motion estimation, Figure 2.14 (f) [Caplovitz et al., 2006]. The high temporal of input events delivered by the DVS sensor leads to motion components that can be considered as to mainly represent motion components tangential to a rotational sweep. However, since those local measures are noisy and need to be integrated over a temporal window, the rotational components become more prominent and gradually deteriorate the IOC solution.

2.7.2 Motion integration based on biologically plausible

Integration strategy based on the IOC mechanism is unlikely to be utilized by the human visual system. In the biological visual system the local motion responses of area V1 are weighted and integrated in cortical area MT. The RF of cells in area MT are larger in their size by up to an order of magnitude. In other words, such cells operate at a much larger spatial context to properly integrate localized responses, similar to the VA method. As a consequence, localized feature responses at line ends or corners lead to stronger responses in the integration process.

In order to account for integrating local motion responses of unknown components and compositions, Chapter 3 introduces an extension of our model by using motion integration algorithm following to the study of [Bayerl and Neumann, 2004] and [Bouecke et al., 2011]. In this extension, we utilize model mechanisms of cortical area MT that integrate the normal flow components derived from the spatio-temporal filter responses defined in Section 2.5.2.

2.8 Discussion

Computational mechanisms of visual processing often utilize conventional frame-based camera for data acquisition. Many works have addressed motion estimation using the frame-based imager, which can be characterized as computer vision approaches (such as [Lucas and Kanade, 1981], [Brox et al., 2004], [Drulea and Nedevschi, 2013]) and bio-inspired related models (such as [Adelson and Bergen, 1985], [Strout et al.,] [Emerson et al., 1992], [Challinor and Mather,

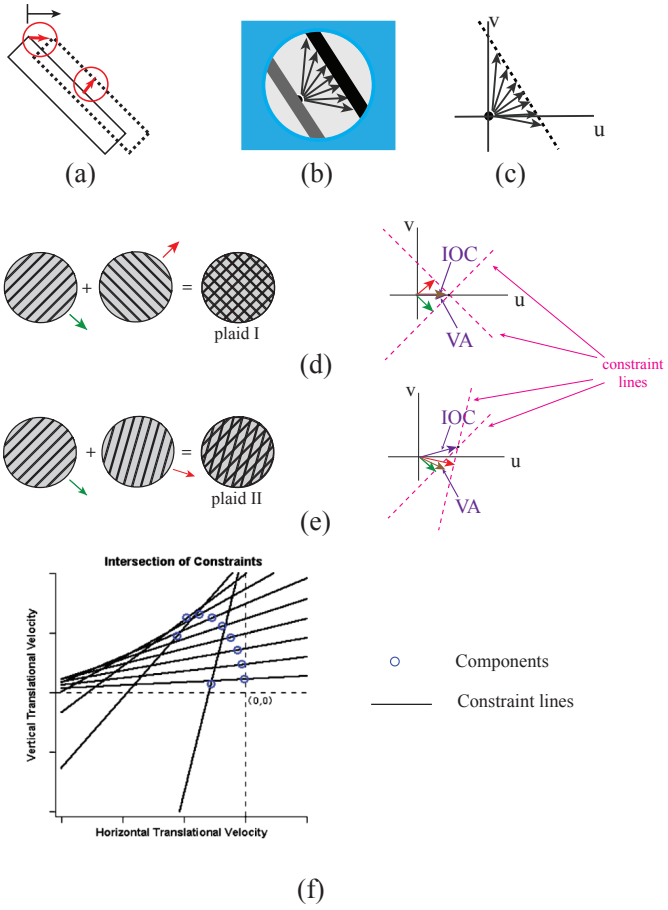


Figure 2.14: Aperture problem. (a) Elongated bar moves to the right direction, the red circles represent the detector cells at the corner and center of the bar. (b) A family of possible motions occurring at a center of bar contour. (c) The family of motions are represented in the velocity space. (d) Plaid I composes by two gratings that are oriented in different directions. Both gratings are moved in directions orthogonally to their contrasts. In this case, the motion directions which are predicted by IOC and AV are approximately similar. (e) Plaid II composes by two gratings that are oriented and moved in the same side with respect to the generated movement of 2D crossings lead to predict different motion pattern by IOC and AV approaches. (f) The IOC approach fails with the rotational motion (this image is taken from [Caplovitz et al., 2006]). The blue circles represent the direction and magnitude of component motion vectors in velocity space for various points along the leading edge of an ellipse undergoing clockwise rotation. The constraint lines represent all the possible real-world motions which could be generated at the contour locations. The IOC approach can be applied to translational motion for which all constraint lines share a common point of intersection. While in the rotational motion, the IOC approach fails to predict the motion because there is no common point of intersection.

]). However, the concept of visual processing based on frames is far from the principles of the biological visual system. Neuromorphic vision sensors is a spike-based vision sensor that simulates the neural circuitry of the biological retina. Recently, [Benosman et al., 2012] and [Tschechne et al., 2014b] carried out motion estimation using retina sensors in which the first article adopted a computer vision approach, while the second considered a bio-inspired model. Along this chapter, we introduced and discussed the available methodologies that concerning motion estimation based on events cloud. We show that the Neuromorphic vision sensor needs an adaptive model for motion estimation that is compatible with address events representation principles. We propose a neural model for motion estimation using Dynamic Vision Sensor (DVS). The neural model processing is inspired by the low-level filtering at the initial stage of the visual system. We adopted the spatio-temporal filtering model suggested in [Adelson and Bergen, 1985] and integrated new temporal filters to fit with AER principles. In order to regulate the overall activation of single neurons against a pool of neighboring cells, we incorporate a competitive stage that operates upon the spatial as well as the feature domain.

Our approach contributed to bio-inspired model for motion estimation using DVS sensors by introducing new temporal filters consistent with polarity responses of the retina sensors. According to [Adelson and Bergen, 1985], the temporal filters in bio-inspired models defined as a smoothing functions with biphasic shape responses, in which temporal gamma functions of different duration were used to achieved temporal smoothing and differentiation. These functions can be approximately decomposed into a convolution of a numerical difference kernel with temporal smoothing filters. Since the representation AER uses the first order temporal derivative (i.e., the discrete events generated based on the *changes* in the input), we thus suggest to employ temporal smoothing filters and convolve them with the input stream of events to obtain scaled versions of temporally smoothed derivatives of the input luminance function.

Our model was tested using different kinds of stimuli. Here, we examined the model by probing it with synthetically generated ground truth stimuli and realistic complex motions, e.g. biological motions and a bouncing ball. In many cases, the model shows accurate results for translatory motion estimation compared with synthetic ground truth. The error value increased in rotational motion cases due to the limited number of estimated directions compared with the ground truth. This drawback could be overcome by increasing the estimated directions of our model. The spatial low resolution of the DVS sensor has unfavorable impact on several locations of the star image case due to the spatial aliasing problem which leads to spurious estimations.

The size of the temporal sampling interval can affect the motion estimation results in which smaller temporal window size gives better estimation. This smaller window can capture more motion details since it accumulates the occurred events immediately and transcribes their motions instantaneously. As a consequence, the subtle information can be maintained.

Following the theoretical studies (e.g., [Brosch and Neumann, 2014]) we carried out the normalization process using model neuronal activation which is described by [Grossberg, 1988] as gradual changes of the membrane potential. Activity normalization of model areas V1 and MT is computed by realizing a slightly simplified version of the scheme described in [Brosch and Neumann, 2014] and solved the normalization interaction at equilibrium, namely evaluating the state response for $\frac{dv(t)}{dt} = 0$. Here, the normalization process is carried out in spatial domain by calculating the pool of activities of individual neuron over a circular spatial neighborhood. In addition, we normalized neuron activities in the feature domain. Since area V1 derives directions and spatio-temporal selectivity, we normalized responses of area V1 by averaging activities of each cell over all directions.

The aperture problem can be reduced by integrating the local motion estimation of area V1 in model area MT. The RFs of cells in area MT are operate at a much larger spatial context to properly integrate the localized responses that generated from area V1. This integration is similar to the principles of VA method (see Section 2.7). As a consequence, localized feature responses at line terminators (such as bar ends or corners) lead to stronger responses in the integration process. This integration mechanism has been investigated in the upcoming chapter (Chapter 3) in which a neural model of area MT is introduced and discussed.

Chapter 3

Model Extension: Neural Mechanisms of Cortical Motion Computation

3.1 Introduction

Motion perception is a significant source of information for the cortical visual system in which this information is analyzed and processed by a large number of interconnected areas [Felleman and Van Essen, 1991]. Each area is characterized by cells of different feature selectivity and varying tuning properties [Lamme et al., 1998]. The neuron responses increase in complexity from the lower level up to the higher cortical area [Perrone, 2004]; [Ungerleider and Haxby, 1994]. Motion analysis starts in the primary visual cortex V1 [Chalupa and Werner, 2004]. The responses of area V1 are forwarded to subsequent areas, namely middle temporal area (MT), medial superior temporal area (MST) and beyond [Simoncelli and Heeger, 1998]. Visual areas interact via feedforward and feedback sweep processing (see [Felleman and Van Essen, 1991]; [Bastos et al., 2014]; [Friston and Büchel, 2000]). The feedforward sweep processes information that are transferred from lower to higher levels via bottom-up connections. These connections are paralleled by top-down connections to project the feedback signals from the higher to the lower level via a reverse counter stream network [Felleman and Van Essen, 1991]; [Koch and Davis, 1994]. Feedback processing tends to act as a modulator to amplify lower stages activations [Larkum, 2013].

Although neuroscientists confirm the influence of bidirectional integration among cortical visual areas [Felleman and Van Essen, 1991]; [Johnson and Burkhalter, 1997]; [König and Luksch, 1998]; [König and Luksch, 1989], the

role of feedback processing is still not fully uncovered [Crouzet and Cauchoix, 2011]. The question is how the higher levels areas can project activations along feedback stream to the lower areas in which the response of each area is characterized by specific properties? In this chapter, we propose a possible solution to this question by suggesting a mechanism of feedforward and modulatory feedback interaction. Here, we extend our suggested model in Chapter 2 by adding a brain function, namely middle temporal (MT). The extended model simulates the visual areas interactions by proposing a new mechanism of feedforward and feedback processing between two main model visual areas with different feature selectivity, V1 (direction selectivity) and MT (velocity selectivity). The initial motion representation generated in model area V1 is weighted and is then subsequently pooled over a local neighborhood defined in the spatio-temporal domain in model MT using cells with larger receptive fields (RFs). Here, a new model of MT motion selective neurons which are tuned to different speeds and directions is introduced.

To enhance the initial motion estimation in area V1 and reduce ambiguity in the estimated motion, top-down feedback signals from area MT are fed back and re-entered in area V1. Along the feedforward path different feature representations of the spatio-temporal input are generated through hierarchically organized filter processes. The specific features extracted are spatio-temporal frequency (in V1) and velocity (direction, speed) at spatial locations (in MT), respectively. The model specifically suggests how MT feature representations generate proper feedback signals that integrate with the spatio-temporal responses in V1. In a nutshell, the re-entry is accomplished through an integration of MT activations over different speed selectivities along a preferred direction. Coarse-grained activations integrated over neighboring direction selectivities emphasis corresponding spatio-temporal directions via modulatory feedback. In the model, neurons activations of both areas, V1 and MT, against the pool of neighboring neurons are carried out to adjust the neurons activities in the spatial as well as in the feature domains. In the following Sections the details of our extended model are introduced.

3.2 Motion processing architecture

The architecture of the neural model is inspired by the mammalian visual system in which the input data is evaluated and processed by a hierarchical model of different brain areas. Figure 3.1 shows an overview of the feedforward and feedback neural model structure. The brain areas interact through bidirectional feedforward and feedback signal pathways. Motion analysis starts in the primary visual cortex (V1) in which the responses of V1 cells are for-

warded to subsequent areas in the model, particularly the middle temporal area (MT) which directly receives V1 projections (see [Crouzet and Cauchoix, 2011]; [Koivisto et al., 2011]). In turn, area V1 receives re-entrant input from MT neurons through top-down feedback connections. The action of such re-entrant activation is modulatory in its nature as it cannot by itself generate activity at the reentry target location. In addition, the activation in area MT can be modulated through higher level areas such as the medial superior temporal area, MST, (see [Felleman and Van Essen, 1991]). Since we do not incorporate a model MST in the current architecture, this modulatory input is left void. Figure 3.2 shows the hierarchical architecture of V1-MT feed-forward and feedback processing in which the following subsections detail the processing stages of the model.

3.2.1 Motion estimation-feedforward processing

In our model, the function of the feedforward network architecture is depicted in three stages to show how the processing is implemented dynamically. In the first stage, Figure 3.2(A), the spares of input ON/OFF events are fed to the model mechanism of area V1 in which this stage was introduced in Section 2.5.1. In the second stage, Figure 3.2(B), the model mechanism of area V1 is defined by three column layers which are composed of filtering, modulation and normalization. Cells in this area are selective to movement direction. Due to the small RFs of V1 neurons, local motion can only be detected along direction orthogonal to extended contrasts (i.e., aperture problem). The model mechanism of area V1 including the three column layers was presented in Section 2.5.2. The local estimated motion is, then, weighted and integrated in the third stage (i.e., area MT). Here, cells are modeled using a new mechanism that are selective to different directions and speeds. The following subsections detail the proposed models of area MT.

3.2.1.1 Model mechanism for area MT

In area MT, again, the structural model is defined by three stages, namely filtering, modulation, and normalization, as shown in Figure 3.2(C). In the filtering stage, the incoming visual responses from area V1 are weighted and pooled with larger RFs. Here, we propose a neural model for area MT where neurons are selective to different velocities. In essence, we modeled the neurons in area MT by taking into account the following features:

1. Speed and direction selectivity of MT cells encoded by proper sampling V1 cells responses in space-time: Cells in area V1 which are driven by

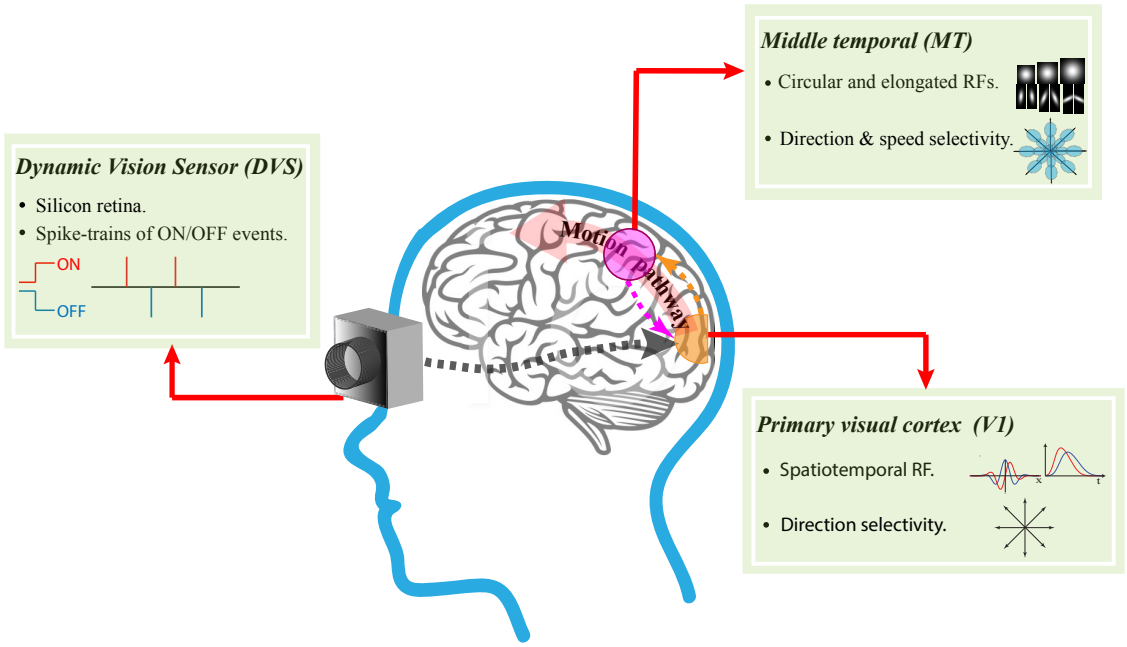


Figure 3.1: Feedforward and feedback neural model structure. The spike of ON/OFF events are fed to the first visual area of motion analysis in the motion pathway. Along feedforward processing the output of area V1 is forwarded to the subsequent area MT. The feedback signals from area MT are re-entered to area V1 through feedback interaction.

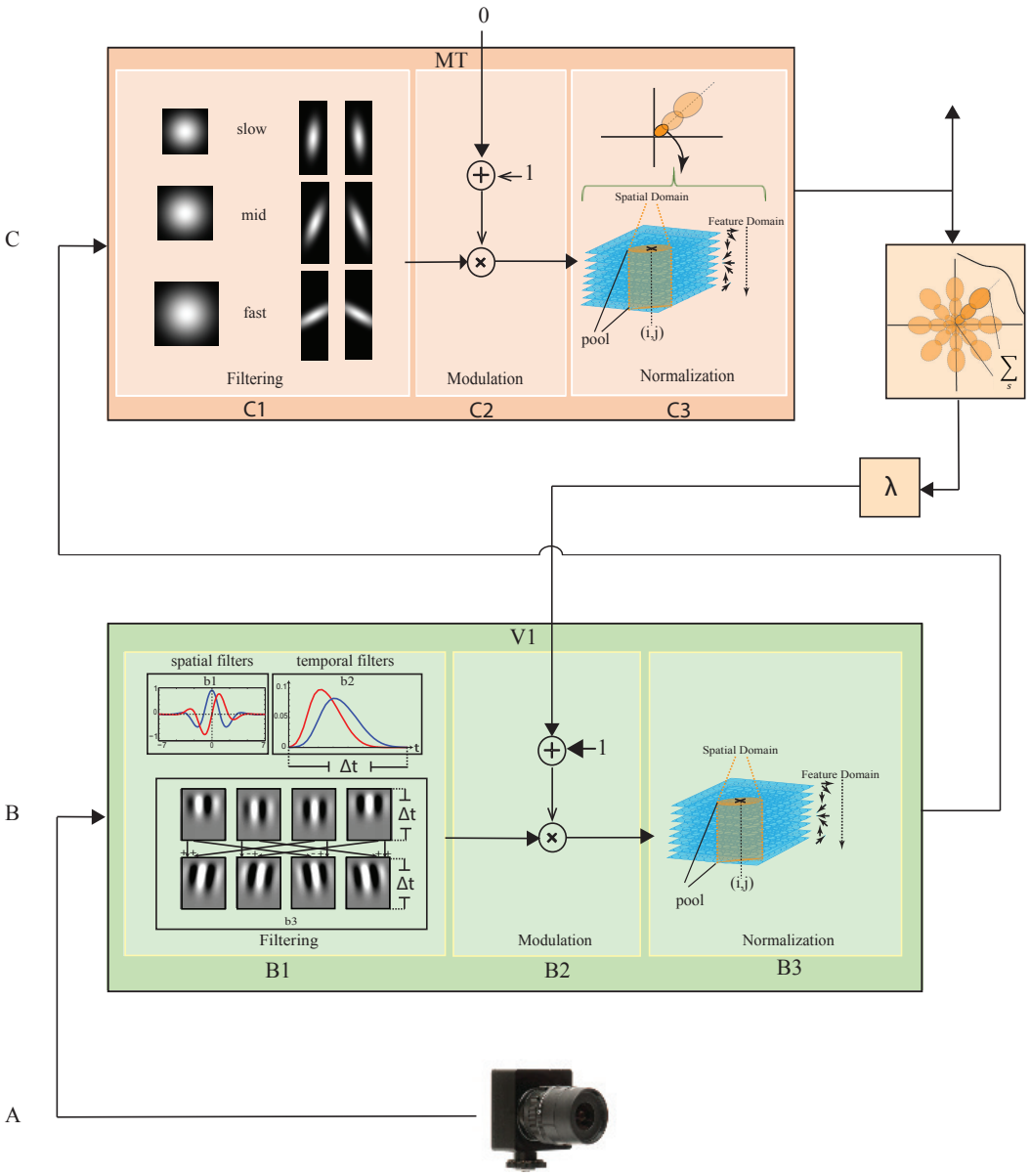


Figure 3.2: Block diagram of V1-MT feedforward and feedback processing. (A) DVS input (see Section 2.5.1 for more details about the DVS camera). (B) The model of area V1 (B1) Spatiotemporal filter construction. (b1) Spatial filters. (b2) Temporal filters. (b3) The first row represents the products of two spatial and two temporal filters; the second row represents the sum and difference of the product filters. (B2) modulation of area V1 based on the feedback of area MT. (B3) normalization mechanism of area V1. (C) The model of area MT. (C1) filtering representation of MT cells. (C2) modulation signal for area MT based on area MST activation which, here, is set to zero. (C3) MT normalizing mechanism.

a sweep of input stimulation are distributed over a larger spatial field of locations. Due to their small RFs such cells respond spatio-temporally rather coarsely, but can distinguish directions (left vs. right orthogonal to the contrast orientation). In area MT, cells have larger RFs that integrate such spots of V1 activations for a given speed-range. Thus, speed representation in MT neurons can be encoded via the V1 cells population with respect to the temporal axis. Figure 3.3 shows rectangular object moving from left to right at different speeds (slow, mid, fast). The translation of the object corresponds to an oblique stream of ON/OFF events in the spatiotemporal (x,t) domain (we omitted the y-component to keep illustrations simple). The slope of the distributed stream of events with respect to the t-axis reflects the speed of the motion. Cells in area V1 are activated according to these distributions of events and thus generate a population of activated cells that are distributed along different orientation in the spatio-temporal domain. The responses of V1 cells are integrated in the space-time domain by MT cells with larger RFs. This kind of integration equips MT cells with contextual information for a particular speed-range.

2. Organization and geometry of MT receptive fields: The RFs geometry of MT neurons has been investigated by [Born, 2000], where the authors identified two types of MT RF profiles: classical receptive field CRF and center-surround cells. CRF responds best to wide field motion, while center-surround interaction are sensitive to motion contrasts, which can be described by center-surround cells. We focus on CRF cells in our model in which weighting functions with different sizes are utilized. These cells are organized in the spatial-temporal domain (x,y,t). We use Gaussian fall-off functions to integrate the spatial responses of V1 cells over larger RFs (V1:MT 1:3, 1:3.5 and 1:4). In addition, elongated weighting filters are used to integrate the responses of V1 cells population with respect to the temporal axis. The integration process of V1 responses over time axis encodes the speed representation in area MT. The elongated weighting filters are tuned to different speeds (fast, mid, and slow) by orienting them in the spatiotemporal domain in which fast motion tuning filters are more oriented with respect to the temporal axis. Following the study of [Grzywacz and Yuille, 1990], we increased the size of the RF profiles with increasing speed selectivity. Figure 3.4 (a) shows the model of MT receptive fields in 2D space (x,y) and (x,t) while Figure 3.4 (b) demonstrates the model of MT receptive fields in 3D space (x,y,t). The responses of such cells are calculated by

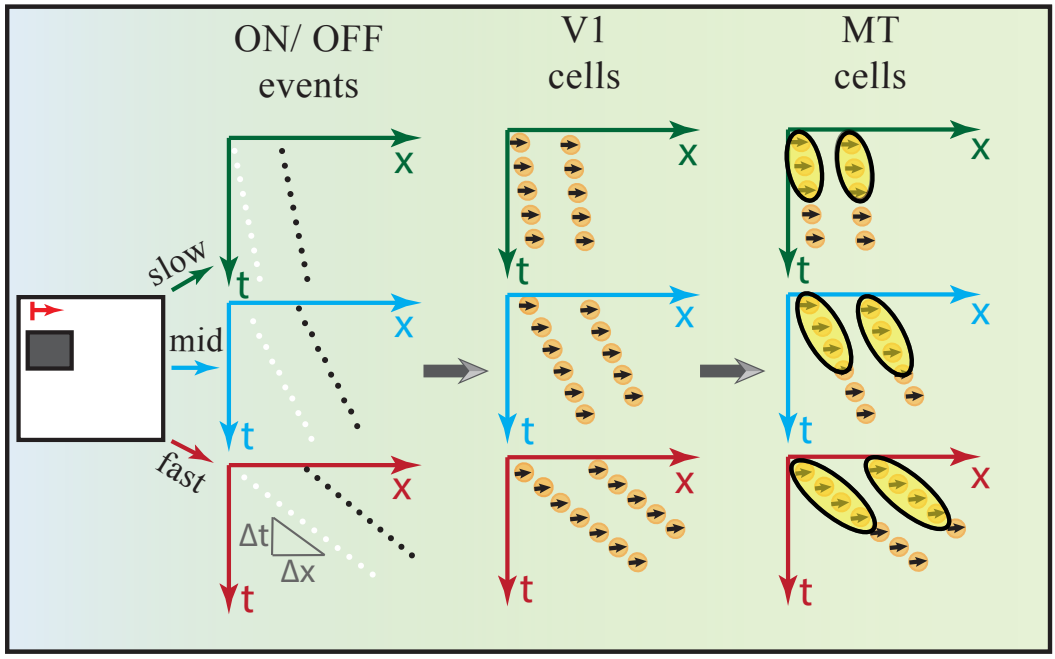


Figure 3.3: speed representation in spatiotemporal (x,t) domain. A rectangular object is moving from left to right at different speeds. In the first column, streams of ON/OFF events are generated in the spatiotemporal domain (x,t) for different speeds (slow, mid, fast). The sequences of oriented in the space-time domain corresponding to the speed of the motion. In the second column, sketch of the activated V1 cells in the 2D spatiotemporal domain driven by a slow, mid and fast sweep of input stimulus. The third column shows how cells in area MT encode the speed of the motion via integrate spots of V1 activations in spatiotemporal domain

$$r^{MT}_{\theta,s_i}(x,y,t) = \sum_{\bar{\phi}} \left\{ r^{v1}_{\theta} * \Lambda_{\sigma_i} * \Psi_{\bar{\phi}_i} \right\}_{x,y,t}, \tag{3.1}$$

$$\Lambda_{\sigma_i} = \frac{1}{2\pi\sigma_i^2} \cdot exp\left(-\frac{x^2+y^2}{2\sigma_i^2} \right), \tag{3.2}$$

$$\Psi_{\bar{\phi}_i} = \frac{1}{2\pi\sigma_x\sigma_t} \cdot exp\left(-\frac{\check{x}^2}{2\sigma_x^2} + \frac{\check{t}^2}{2\sigma_t^2} \right), \tag{3.3}$$

where $\begin{pmatrix} \check{x} \\ \check{t} \end{pmatrix} = \begin{pmatrix} cos\bar{\phi}_i & -sin\bar{\phi}_i \\ sin\bar{\phi}_i & cos\bar{\phi}_i \end{pmatrix} \cdot \begin{pmatrix} x \\ t \end{pmatrix}$, the symbol $'*'$ indicates the convolution operator, Λ_{σ_i} denotes the Gaussian weighting functions in which σ_i is the spatial extent, $\Psi_{\bar{\phi}_i}$ represents the elongated Gaussian functions where $\bar{\phi}_i$ is the orientation of the Gaussian function (ϕ_i, $-\phi_i$ define left and right orientations), r^{v1}_{θ} is the response of area V1 in which θ denotes motion direction selectivity, the index s_i represents speed selectivity $s_i = \{fast, mid, slow\}$ that corresponds to the spatial extent σ_i and the orientation $\bar{\phi}_i$.

3. Temporal integration: We incorporated a temporal trace rule in our model in order to hold an amount of past neural activations in which the weight of the present activities is higher than the past activities. The extent of such temporal windows takes into account the temporal impulse responses of MT cells reported in, e.g., [Rolls and Stringer, 2006]. Such a rule has also been adopted in several studies (e.g., [Layher et al., 2014a]; [Sullivan and de Sa, 2004]).

$$\dot{r}^{MT_{out}}_{s_i,\theta}(t) = -\delta r^{MT_{out}}_{s_i,\theta}(t) + r^{MT_{in}}_{s_i,\theta}(t), \tag{3.4}$$

where δ $(0 < \delta < 1)$ represents the strength of the temporal integration and $\dot{r}^{MT_{out}}_{s_i,\theta}(t)$ denotes the calculated average activity of MT cells, $r^{MT_{out}}_{s_i,\theta}(t)$ and $r^{MT_{in}}_{s_i,\theta}(t)$ represent the input and output activities of the MT cells.

3.3 Response normalization

Experimental investigation have shown that the responses of cells in visual cortical areas show significant nonlinearities depending on spatio-temporal activity distribution in the cell activation in the space-feature domain surround-

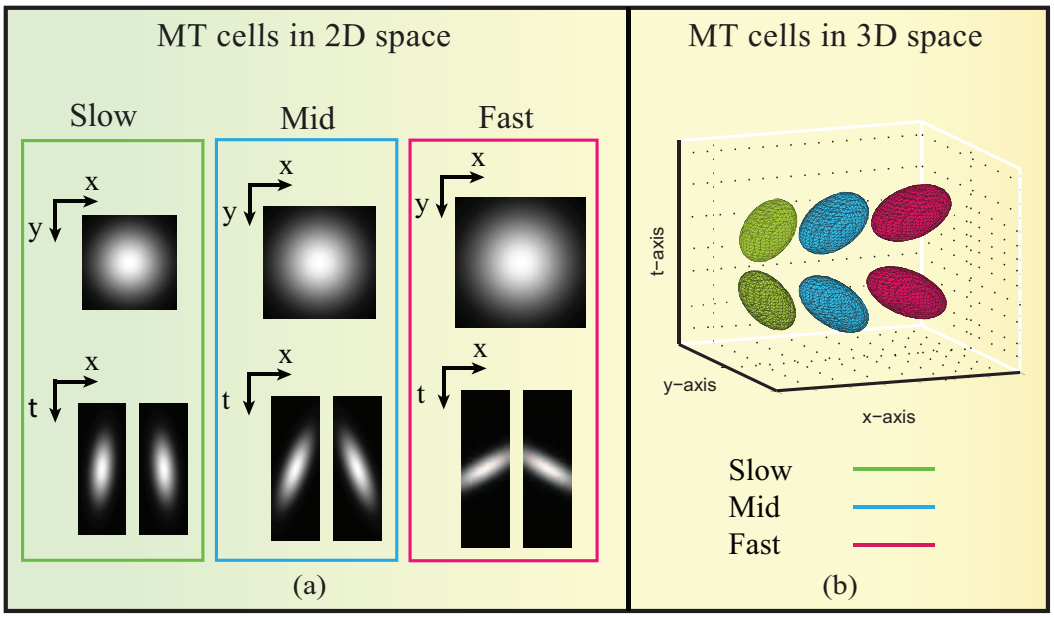


Figure 3.4: The model of area MT. The RFs of cells in area MT are modeled using Gaussian functions with circular and elongated shapes. (a) The model of MT receptive fields in 2D spatial space (x,y) and 2D spatiotemporal space (x,t). (b) The model of MT receptive fields in 3D space (x,y,t).

ing a target cell [Heeger, 1992]; [Carandini and Heeger, 2012]. Such response nonlinearities have been demonstrated in the LGN, early visual cortex (area V1), and beyond. In many studies, it has been proposed that responses are calculated via the normalization of the target cell response based on pooling activities in a neighborhood (see [Brosch and Neumann, 2014]). Model neuronal activation is described by gradual changes of the membrane potential which is here considered as a population state variable, as described by [Grossberg, 1988]

$$\tau \frac{dv\,(t)}{dt} = -A \cdot v\,(t) + (B - C \cdot v\,(t)) \cdot net_{ex} - (D + E \cdot v\,(t)) \cdot net_{in}, \quad (3.5)$$

where A represents the passive leakage, B and D are parameters denoting the saturation potentials (relative to C and E, respectively), and net_{ex} and net_{in} denote generic excitatory and inhibitory inputs to the target cell with the state, or membrane potential. Here, the net inputs are defined by the responses encoding motion from the filtering stage.

A firing-rate function generates positive output activations from the current state level, or membrane potential. Here, we employ a simple rectification to yield $r(t) = max(0, v(t))$. We set the parameters in Equation (3.5) such that the output range for the membrane potential is bounded to the range $[0, B]$, setting B=1 and D=0 and C=E=1.

In order to achieve balanced activities of individual cells against neighborhood activities, we employ a normalization mechanism in which neuron activities are adjusted in the spatial and feature domains. We solved Equation (3.5) at equilibrium state $\frac{dv(t)}{dt} = 0$, using the parametrization for constants A to E as defined above

$$v_\infty = \frac{net_{ex}}{A + net_{ex} + net_{in}}. \quad (3.6)$$

As outlined in Figure 3.2 the activities of areas V1 and MT are normalized in the third stage B3 and C3, of each individual processing cascade, respectively. The normalization of the V1 motion selective responses take the spatial and the feature domain into account. Here, the features considered are orientations and directions. The model responses r^{v1} are normalized in the spatial domain using a Gaussian weighting function defined over the spatial domain. The motion selective responses are defined in orthogonal direction space relative to the local contrast orientation θ of the spatial filter kernels used. We take the direction feature space into account as well by summing activity over all directions and then scaling this sum by N. In all, we can denote the overall

pool activation by

$$r^{v1pool}(x,y) = \frac{1}{N}\sum_{\theta}\left\{r_{\theta}^{v1} * \Lambda_{\sigma_{nor}}\right\}_{x,y}, \tag{3.7}$$

with θ denoting the motion directions, $'*'$ denotes the (spatial) convolution operator, N is the number of contrast filter orientations and Λ is the spatial weighting function of the pooling operation. The latter is a Gaussian function (as defined in Equation (3.2)) which is parametrized by the parameter σ_{nor} to denote the width of the spatial extent. Finally, the resulting normalized responses in the spatial and feature domains is calculated by

$$r_{\theta}^{v1nor}(x,y) = \frac{r_{\theta}^{v1}(x,y)}{A + r_{\theta}^{v1}(x,y) + r^{v1pool}(x,y)}. \tag{3.8}$$

Like in Equation (3.6) constant parameter A refers to the passive decay of the corresponding dynamic mechanism and avoids zero division.

Similarly, the activity of individual neurons in areas MT are balanced in the spatial domain as well as in the feature domain. Here, features are defined in the velocity space (both directions and speed). In the spatial domain, again, we used a fall-off function $\Lambda_{\sigma_{nor}}$ to weight the activity in the spatial neighborhood of the target cell according to the width of the spatial extent σ_{nor}. While in the feature domain, the motion selective responses of MT neurons are defined in velocity space (direction and speed). We adjust cell activations against the pool of neighboring cells in the direction domain at each particular speed. This is accomplished by averaging activities of each cell over all directions $\{\theta_k\}_{k=1\cdots N}$ of each speed s_i. The pooling activities of individual neuron over spatial and feature domain are defined by

$$r_{s_i}^{MTpool}(x,y) = \frac{1}{N}\sum_{\theta}\left\{r_{\theta,s_i}^{MT} * \Lambda_{\sigma_{nor}}\right\}_{x,y}, \tag{3.9}$$

where the index s_i represents the speeds, $i = \{1 \cdots 3\}$ corresponds to (slow, mid and fast). The resulting normalized response of neurons in area MT in the spatial and feature domain is finally calculated by

$$r_{\theta,s_i}^{MTnor}(x,y) = \frac{r_{\theta,s_i}^{MT}(x,y)}{A + r_{\theta,s_i}^{MT}(x,y) + r_{s_i}^{MTpool}(x,y)}. \tag{3.10}$$

Again constant parameter A refers to the passive decay of the corresponding dynamic mechanism and avoids zero division.

3.4 Feedback from area MT to V1

Along the feedforward processing sweep the feature representations in area V1 and MT, respectively, are different. Cells in area V1 are direction-selective while cells in area MT are velocity-selective (direction and speed). In other words, the bottom-up signals (V1→MT) predominantly drive from direction features (and do not distinguish between different speeds), while the top-down signals (MT→V1) drive from both direction and speed features. The question here is how can the higher area signals interact with the lower area where both are fed by different feature domains? The model introduces a possible solution to this question by suggesting a mechanism of modulatory feedback interaction. In this mechanism, we propose integrating the top-down responses over different speeds along a particular preferred direction. This integration transforms the feature responses of the MT cells from direction and speed domains to the direction domain. In other words, integration process over different speeds (slow, mid and fast) along each direction generates feedback signals derived from representations sensitive to motion directions rather than directions and speeds. Here, the feedback signals from area MT can be re-entered at the lower area V1, in which the properties of both responses (feedforward and feedback) are matched. Area V1 cells receive feedback signals from the higher area MT such that the feedforward activities can be modified via the modulatory influence of feedback signals. The top-down signals alone, however, cannot provoke any activities when feedforward signals are absent. We utilized the feedback modulatory mechanism suggested by [Bouecke et al., 2011]; [Neumann and Sepp, 1999], which in abstract terms is denoted by,

$$r^{out}_{x,y,feat} \propto r^{FF}_{x,y,feat} \cdot (1 + \lambda \cdot r^{FB}_{x,y,feat}). \tag{3.11}$$

In this scheme FF denotes to the feedforward signal stream and FB refers to the feedback stream. The index $feat$ defines the considered feature in both feedforward and feedback stream which is here represents motion direction at the spatial position (x, y), and λ denotes a constant amplification factor. In order to keep the smoothness of the top-down signals in the feature-space, we weighted the activities in MT over the neighborhood directions using a Gaussian function Λ_{σ_m}. The broad of the weighting activities over the neighborhood directions is parametrized by σ_m. The modulated V1 responses in our model are finally calculated by

$$r^{v1_{mod}}_{x,y,\theta} = r^{V1}_{x,y,\theta} \cdot (1 + \lambda \cdot \sum_{s_i} r^{MT_{nor}}_{x,y,\theta,s_i} \cdot \Lambda_{\sigma_m}), \tag{3.12}$$

where Λ_{σ_m} is Gaussian weighting function of the smoothing operation over

direction θ in which σ_m denotes the width of the smoothing extent, s_i refers to discrete speed ranges, indexed by $i = \{1 \cdots 3\}$ and λ defines the strength of the feedback projection from area MT. In the case of a pure feedforward processing sweep, the modulator feedback signals are switched off by setting $\lambda = 0$. As a consequence, the filtering response signals in model area V1 are forwarded to the third column stage in which the responses of V1 cells are normalized in the spatial and feature domains.

3.5 Results

To demonstrate the potential of our model, we used synthetic ground truth data along with a set of different stimuli with translator and rotational motions. The motions of the stimuli were recorded using the dynamic visual sensor (DVS 128 sensor) that introduced by [Lichtsteiner et al., 2008]. In our model, the size of the spatial filters in area V1 has been defined as a function of the spatial frequency [Watson and Ahumada, 1983] and [Watson and Ahumada, 1985] such that $\sigma_s = \frac{0.5622}{f_s}$. This parametrization is taken in accordance to the suggestion in [Escobar and Kornprobst, 2012]. We have used a spatial frequency of $f_s = 0.25$ in which the size of the spatial kernel in area V1 is (15×15). The estimated motion of our result is based on 8 directions $\theta = \{0°,45°,90°,135°,180°,225°,270°,315°\}$. In the filtering stage of area MT the sizes of the fall-off kernels in the spatial (x,y) domain are (45×45), (49×49) and (53×53) pixels for the slow, mid and fast motion, respectively. These sizes keep the integration process in area MT with larger RFs (V1:MT 1:3, 1:3.5 and 1:4). The sizes of the elongated Gaussian filters in spatiotemporal (x,t) domain are (21×57), (21×63) and (21×69) which are oriented by (240°(left), 300°(right)), (280°(left), 340°(right)) and (175°(left), 355°(right)) for slow, mid and fast motion respectively. In order to hold an amount of past neural activation, the strength of the temporal integration δ in Equation (3.4) has been set to 0.5. In the normalization stage of areas V1 and MT, we set the standard deviation σ_{nor} of the spatial weighing function to 15 which exceeds the size of the RFs of both areas. Here, the passive decay parameter is set to a small value ($A = 0.01$) to avoid zero division. In the feedback processing stream we set λ to 0.8 and σ_m to 2. The following subsections introduce our experimental results to probe the speed sensitivity in area MT. In addition, motion representation in areas V1 and MT via feedforward processing stream will be demonstrated along with the motion representation of the modulated V1 via feedback processing stream.

3.5.1 Speed selectivity in areas V1 and MT

To verify the speed sensitivity of the MT cells, a translating stimulus with different speeds was used. The stimulus movements were recorded using DVS 128 sensor that was mounted on a tripod and placed 25 cm away from the center of the stimulus. We used a dark bar stimulus ($5 \times 95mm$) moving from left to right on a light background. The bar was moved with different speeds (slow 7 cm/sec, medium 17 cm/sec and fast 23 cm/sec) using a linear actuator. Figure 3.5 shows the movement of the bar stimulus to the right. The events configuration that are generated by the DVS sensor is shown in Figure 3.5(b) in which ON-events are generated at the leading contrast edge of the bar and OFF-events at the tail edge. Figure 3.5($c_1 - c_3$) show a translating one-dimensional signal of the bar movement in the spatiotemporal (x,t) domain in response to slow, mid and fast bar motion, respectively. The bar translation traces out a diagonal path in the spatiotemporal (x,t) domain in which the slope of the diagonal path reflects the speed of the bar movement. Cells in area V1 are driven by the input stimulus and respond in a spatiotemporal fashion. Figure 3.6 shows the distribution of the V1 responses which are derived from the three speeds (slow, mid and fast) of the bar movements. In the filtering stage of area V1, the spatiotemporal filters generate a confidence for θ fundamental directions. In order to maintain a single response hypothesis, we generate a sum of the V1 responses over the fundamental directions ($\sum_\theta r^{V1nor}_{x,y,\theta}$). This sum is used to present the distribution of V1 responses shown in Figure 3.6. The results reveal that the mid speed of bar movement encodes higher responses. This indicates the response of filtering stage in model area V1 is sensible to a limited speed inside a spatiotemporal frequency bandwidth which is defined by the filter weighting kernel. V1 responses are integrated in (x,y,t) space via area MT in which sizes of the RFs, or kernels, in area MT are larger than in area V1 (V1:MT 1:3, 1:3.5 and 1:4). Here, cells in area MT are modeled using elongated filters which are tuned to different speeds (slow, medium and fast motion) by oriented these filters with 300°, 340° and 355°, respectively. Hence, higher responses are generated when the RFs of MT cells match the diagonal path of motion configurations. Figure 3.7 demonstrates how MT cells integrate V1 responses from spatial locations along the diagonally arranged event configurations (generated by slow, mid and fast speeds of motion). The sketch outlines the initial events and the corresponding localized V1 cell responses for rightward motions occurring with different speeds. The result reveals that cells in area MT that are oriented by 300° are highly responsive to the slow motion whereas the responses of the other cells are greatly reduced with the time, see third column of Figure 3.5($c_1 - c_3$). This is due to the fact that the two cells with higher speed selectivity (orientation

340° and 355°) are activated at the beginning and then these activations diminish over time since the RFs do not match the locations of the diagonal path of motion configurations with the cells preferred tuning. Similarly, filters that are oriented by 340° and 355° are highly responsive to the medium and fast motions respectively. The results confirm that the proposed cells respond correctly to the used speeds (slow, mid and fast).

3.5.2 Motion representation in feedforward and feedback sweeps

A set of different stimuli were recorded using the DVS 128 sensor to evaluate motion representations in the feedforward and feedback sweeps. The DVS sensor is mounted on a tripod and placed 23 cm away from the center of the stimuli. These stimuli were moved via translatory and rotatory motions in which the movements of the stimuli are highlighted in the top-left of the stimuli images. To produce these motions, we used linear and rotational actuators in which the speed of the linear actuator is 20 cm/sec while the speed of the rotational actuator is 5.23 rad/sec. In order to keep a single motion hypothesis, the estimated results of the optic flow in area V1 are generated based on a weighted sum of the fundamental directions θ, which generates a confidence for the motion direction $\left(u_e^{V1}(x,y)\ v_e^{V1}(x,y)\right)^T = \sum_\theta r_{x,y,\theta}^{V1nor} \cdot (cos\,\theta\ ,\ -sin\,\theta)^T$. Similarly, the estimated results of the optic flow in area MT are generated based on a weighted sum of the fundamental directions θ and speeds s_i which can be described by $\left(u_e^{MT}(x,y)\ v_e^{MT}(x,y)\right)^T = \sum_\theta \sum_{s_i}\ r_{x,y,\theta,s_i}^{MTnor} \cdot (cos\,\theta\ ,\ -sin\,\theta)^T$. In order to measure the accuracy of our approach, we calculated the angular error $\Phi^{V1,MT}, (x,y) = cos^{-1}(\mathbf{v}_e^{V1,MT}(x,y) \cdot \mathbf{v}_g(x,y))/(|\mathbf{v}_e^{V1,MT}(x,y)||\mathbf{v}_g(x,y)|)$, where $\mathbf{v}_e^{V1,MT}(x,y)^T = (u_e^{V1,MT}(x,y), v_e^{V1,MT}(x,y))$ and $\mathbf{v}_g(x,y)^T = (u_g(x,y), v_g(x,y))$ represent the estimated motion of the visual areas (V1, MT) and ground truth flow vectors, respectively. The error values in the range of $[0°, 180°]$ are displayed as a histogram.

3.5.2.1 Feedforward motion representation

In the feedforward sweep, motion is estimated via areas V1 and MT in which the normal flow that is derived from the spatio-temporal filter is weighted and integrated using the model filter mechanisms suggested for area MT. In the translatory motion, Figure 3.8, we used two stimuli, namely tiger and ball. The first stimulus, tiger, is characterized with a texture that comprises different striped patterns while the second stimulus, the ball, contains different slanted bars that are connected to form the ball interior. The nature of these

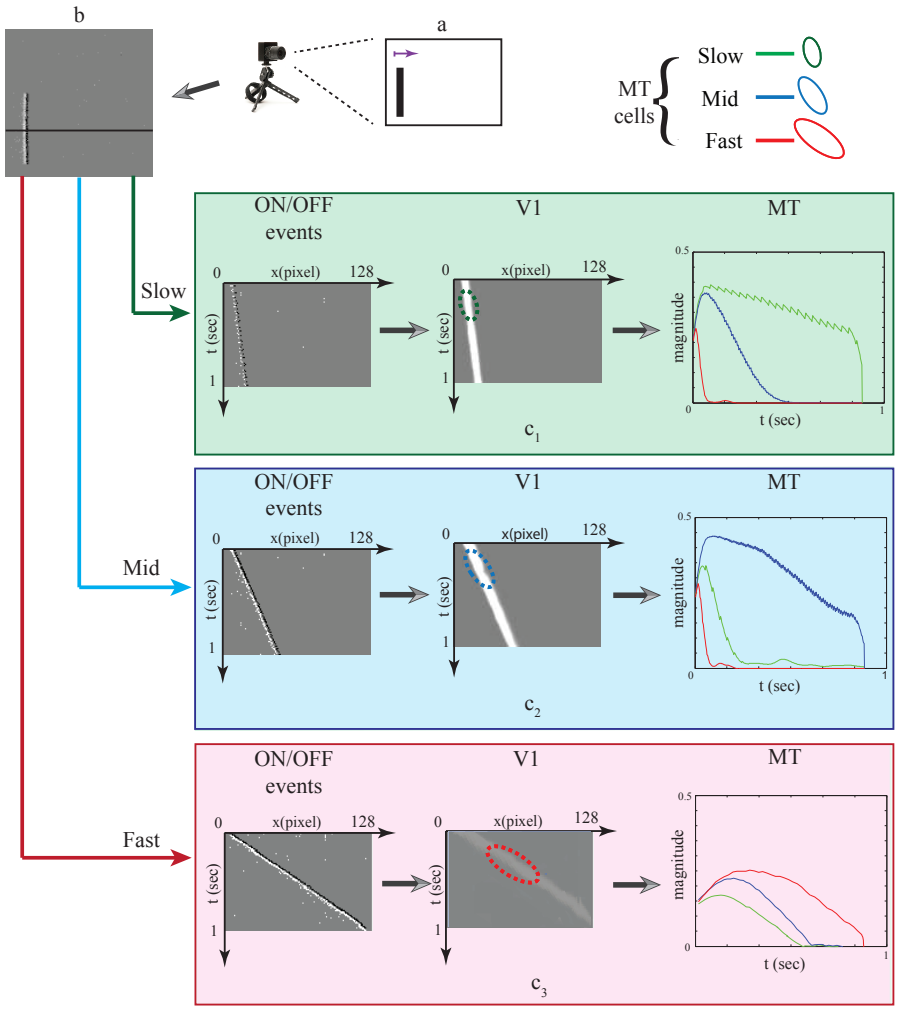


Figure 3.5: Bar movement with three different speeds. The bar was moved to the right with three speeds (slow 7 cm/sec, medium 17 cm/sec and fast 23 cm/sec). (a) Shows the input stimulus. (b) The representation of the ON (white points) and OFF (black points) spike responses that are generated via the DVS 128 sensor. The three blocks (c_1, c_2 and c_3) contain: (left image) the distribution of ON/OFF events in response to a specific speed, (middle image) the V1 cells responses and (right image) the responses of MT cells. The magnitude of V1 cells responses (middle image) are depicted as a brightness in which higher amplitude is generated in response to mid motion while lower amplitude is generated in response to fast motion (see also Figure 3.6). (c_1) The ON/OFF events in the spatial-temporal domain of the bar movement to the rightward with slow motion. (c_2) and (c_3) show the spatial-temporal domain of the bar movement to the rightward with mid and fast motions, respectively. The responses of slow, mid and fast selective MT cells are depicted by solid-green, solid-blue and solid-red lines, respectively. The slow, mid and fast selective cells are depicted over V1 responses as dashed-green, dashed-blue and dashed-red, respectively.

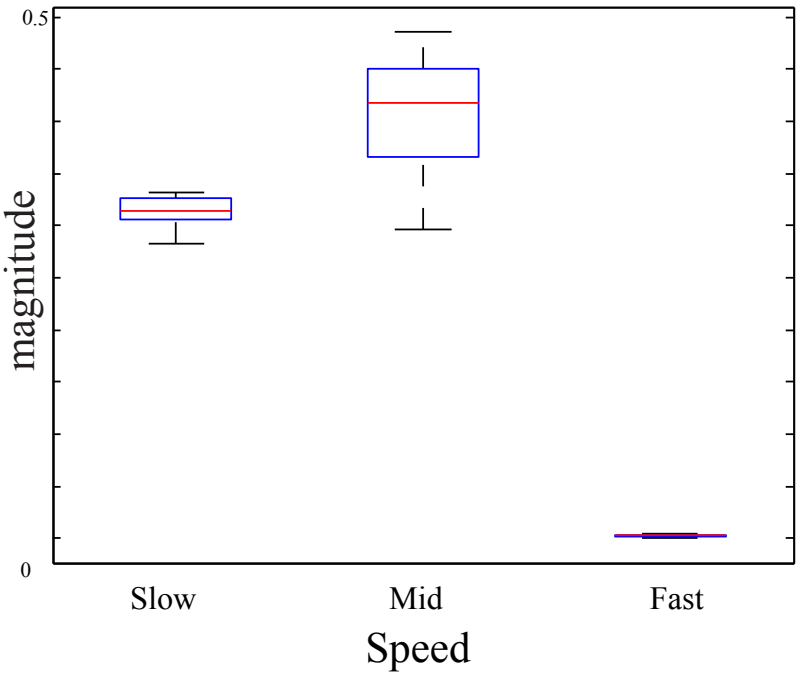


Figure 3.6: The distribution of V1 responses drive by different motion speeds (slow, mid and fast). The box plots display the distribution of the output (magnitude) that are generated by cells in area V1 in response to a specific speed. The first box shows the distribution of the responses in area V1 which are driven by slow bar motion (slow 7 cm/sec). The second box shows the output distribution of area V1 cells in response to mid bar motion (17 cm/sec). The third box illustrates the distribution of the V1 output in response to fast bar motion (23 cm/sec). It can be observed that the mid speed of bar movement encodes higher responses while the fast speed of bar movement encodes lower responses. This indicates that the response of filtering stage in model area V1 is sensible to a limited speed (mid speed) inside a spatiotemporal frequency bandwidth which is defined by the filter weighting kernel.

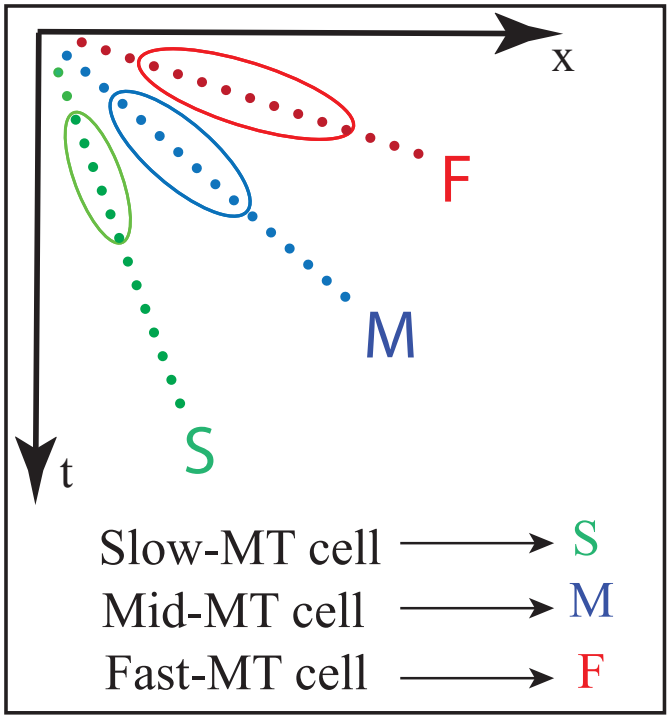


Figure 3.7: MT cell selectivity in the spatiotemporal (x, t) domain. Slow, mid and fast motion selective cells for rightward motion are depicted in green, blue and red, respectively. The diagonal lines of green, blue and red dots represent idealized event responses for slow, mid and fast input motions (with initial inputs generated from the DVS sensor). In accordance to the representation of spatio-temporal inputs increases in speed coincides with an increase in angle relative to the time axis. Model MT cells are suggested to have larger receptive field size in space in comparison to the spatio-temporally selective cells in V1. They also integrate input responses from V1 cells over a temporal period. MT cells with different speed selectivities preferentially integrate V1 responses at the proper spatial offset positions (as depicted in the elliptic outlines). The same representation occurs for leftward motions.

stimuli makes the estimation complex and, then, challenging in translatory motion. The preferred direction of areas V1 and MT are depicted in polar diagrams. In tiger stimulus, the polar diagram shows that the V1 population is selective to a broad range of directions around the correct direction (0°), while this region is shrunk close to (0°) in area MT. In the ball stimulus, motion representation in model area V1 suffers from the aperture problem. Here, the motion is estimated orthogonal to the bar contrast (along bar contour) while the real motion is estimated at the corner regions. Motion ambiguity (aperture problem) is reduced in model area MT due to integrate the unambiguous motion (real motion) with larger RFs. The polar diagram of the direction selectivity shows that the responses of area MT are more tuned towards (0°) than area V1. According to the error-histograms of both stimuli, the error values are reduced and concentrated in the range of [0°, 15°) which indicates that motion is estimated more accurately in area MT compared with area V1. These results are consistent with physiological findings (see e.g., [Churchland et al., 2005]) which showed that cells in area MT are highly directionally-selective compared to cells in area V1.

Neurons in the primary visual cortex area V1 that are selective to spatio-temporal stimulus features have small RFs, or filter sizes. Consequently, they can only detect local motion components that occur within their RFs. That means along elongated contrasts only ambiguous motion information can be detected locally. It is the normal flow component that can be measured along the local contrast gradient of the luminance function (aperture problem). To probe the performance of our model to reduce motion ambiguity in area V1, we recorded a stimulus namely temp·1. This stimulus contains a set of black bars that are slanted with 45° and moved in a direction that differs from the normal flow as highlighted in the top-left of the stimulus. The drifting bars is seen through a circular aperture as shown in Figure 3.9. The result reveals that the V1 population responds in a direction (135°) which referred to the motion was estimated as orthogonal to the bar contrast. This is because the motion was estimated in a local surround, nevertheless actual motions were estimated at the bar ends. In other words, the normal flow can be computed along the local one-dimensional contour while at the ends of the bars the local two-dimensional structure enables to compute the actual motion direction. In area MT, cells integrate initial responses of model area V1 in which the sizes of the RFs are larger (V1:MT 1:3, 1:3.5 and 1:4). Such cells operate at a much larger spatial context to properly integrate localized responses. As a consequence, localized feature responses at line ends lead to stronger responses in the integration process. In all, this leads to that the preferred direction of area MT is more tuned toward the correct direction (90°). The error-histogram

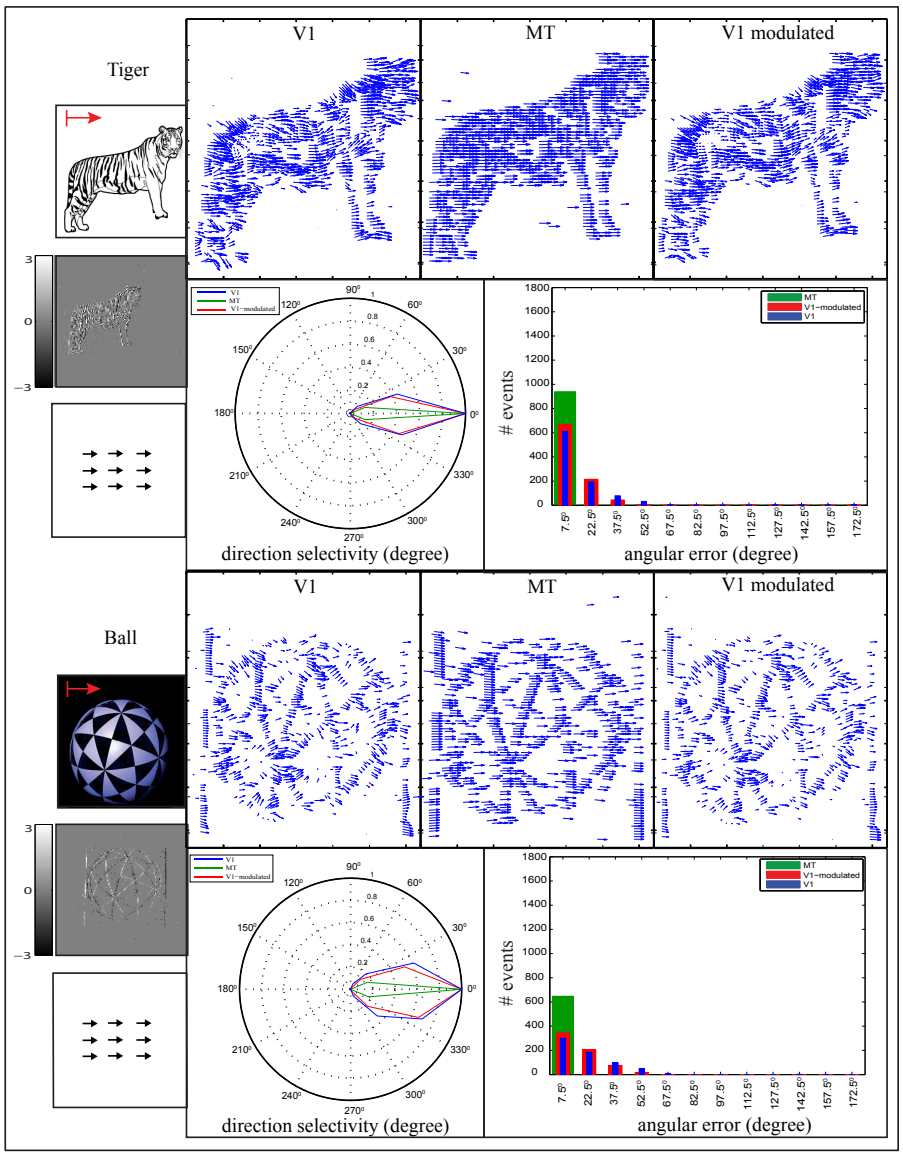


Figure 3.8: Motion estimation of tiger and ball stimuli. The stimuli are moved in rightward direction $(0°)$*. The first column of each stimulus contains the input image, accumulated events* e^{on} *and* e^{off}*, and a sketch of the ground truth optical flow field. The first rows of the second, third and fourth columns represent the estimated motion in areas V1, MT and the modulated V1, respectively. The direction selectivity for these areas are depicted in the polar plot where blue, green and red lines represent the responses of V1, MT and V1 modulated by MT, respectively. The histogram shows the angular error between the estimated motion and the ground truth of rightward motion direction. The abscissa of the histogram represents the binning in the range of the angular error* Φ *which are combined into one bar* $[\theta - 7.5°, \theta + 7.5°)$*, and the ordinate represents the number of events. The ball image is adopted from [Azimuth, 2016]*

in Figure 3.9 shows that motion ambiguity is reduced in area MT in which the fourth bin of the error values $[45°, 60°)$ is decreased.

Object terminators, e.g. end-bar or corner, contain unambiguous motion (real motion direction), hence, the motion direction of an object can be enhanced with increasing the number of object terminators. In our test scenarios this has been investigated by recording two other stimuli, namely temp·2 and temp·3 as shown in Figures 3.10 and 3.11 respectively. In the temp·2 stimulus, the slanted bares (45°) are moving through rectangular aperture rather than circular aperture as shown in Fig 3.10. Here, the bars-ends are distributed along the vertical edges of the rectangular aperture allowing more localized real motion to be integrated in area MT. The results reveal that in area V1 the motion was estimated with preferred direction of (135°) (i.e., the motion was estimated as orthogonal to the bar contrast (normal flow)) while the preferred direction of the cells in area MT was shifted to (112°). The error-histogram shows that the ambiguous motion information (angular error $\geq 45°$) is reduced in the responses of area MT compared with the initial estimation of V1 area in which the fourth bin of the error values $[45°, 60°)$ is decreased and sandwiched in the bins of ($< 45°$). According to our results, the responses of area MT which are influenced by the rectangular aperture are more tuned toward the real motion comparing with the responses of MT cells in circular aperture. The rectangular aperture contains larger number of terminators (bar ends) that align vertically on the longer edges, which in turn leads to integrate more unambiguous information through large RFs of MT cells. In contrast, the integration process for unambiguous motion is decreased in the circular aperture due to the terminators are distributed along circular edge. The results confirm that the shape of the aperture and the number of terminators play a key rule in determining the direction of object movement. These findings are consistent with the previous studies (see e.g., [Wallach, 1976]; [Hildreth and Koch, 1987]) that showed that the direction of the perceived motion is biased towards the higher number of the terminators.

Figure 3.11 shows an example of 2D structure of corner terminators, temp·3. This stimulus contains two superimposed gratings where each grating formed by parallel lines. The stimulus was moved in upward direction which is different from the normal flow of both gratings. Here, the perceived motion is not ambiguous anymore in which the responses of both areas V1 and MT are tuned towards the real motion direction (90°). However, in area V1 ambiguous motions are estimated along bars contours (45°) and (135°) while this ambiguity is reduced in area MT. This is due to the existence of the 2D structure of the corners in which real motion can be estimated. The larger RFs of the cells in area MT allow to integrate more real motion direction that are gen-

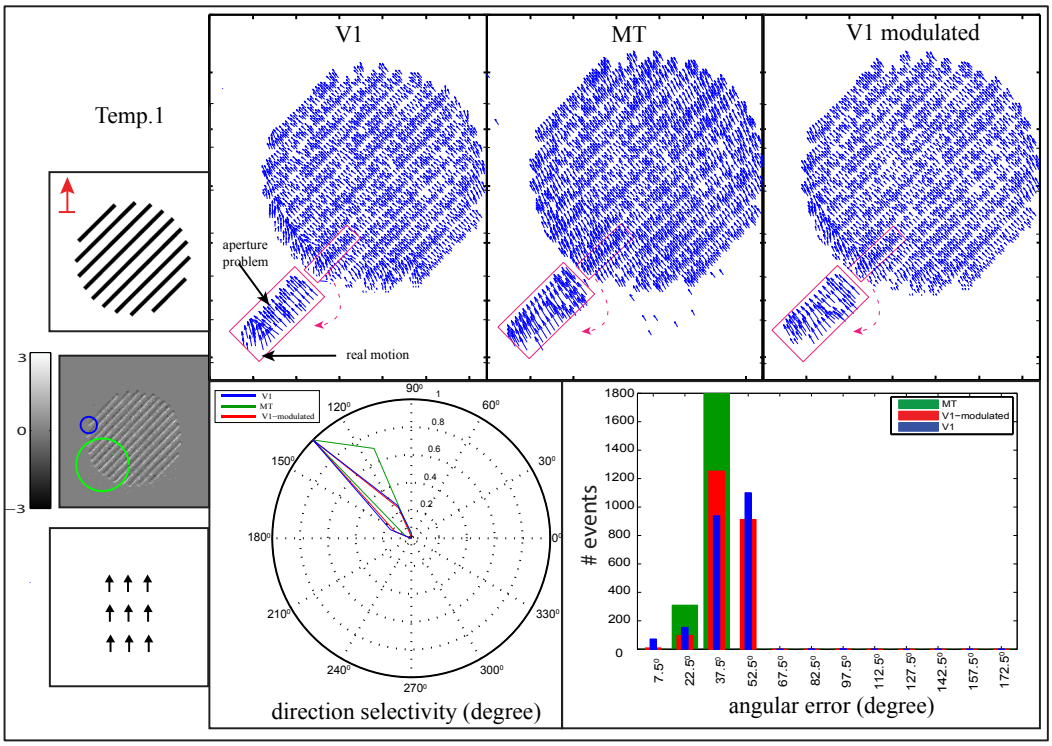


Figure 3.9: Motion estimation of bars movement through circular aperture, temp·1. The bars are oriented by $(45°)$ *and moved in upward direction* $(90°)$ *through circular aperture. The V1 cell and one of the MT cells (V1:MT 1:3) are depicted over the accumulated events as blue circle and green circle, respectively. Motion representation in areas V1, MT and the modulated V1 for each stimulus are shown in the second, third and fourth column respectively. In this stimulus, the real motion direction* $(90°)$ *is estimated at bar endings while the normal flow* $(135°)$ *is estimated along bar contour. The polar plot shows the direction selectivity of V1, MT and the modulated V1 responses which are depicted by blue, green and red lines, respectively. The histogram shows the angular error between the estimated motion and the ground truth of upward motion direction. Here, the ambiguous motion information is reduced in area MT and the modulated V1 in which the angular error between the ground truth motion* $(90°)$ *and the estimated motions is decreased as depicted in the forth bin* $[45°, 60°)$*. The abscissa of the histogram represents the binning in the range of the angular error* Φ *which are combined into one bar* $[\theta - 7.5°, \theta + 7.5°)$*, and the ordinate represents the number of events.*

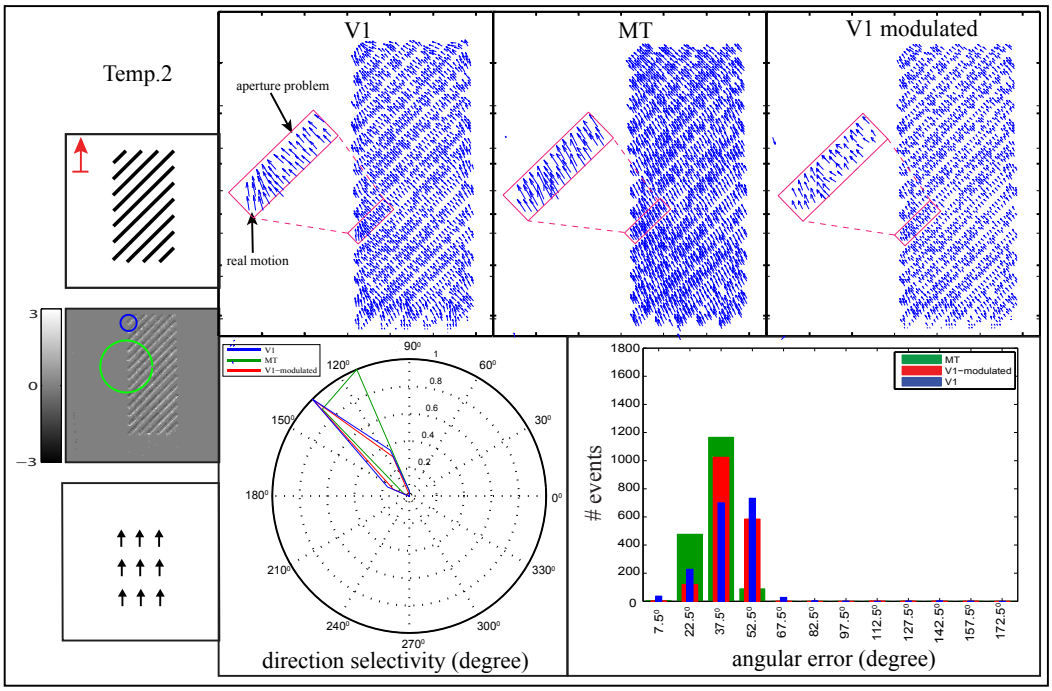


Figure 3.10: Motion estimation of bars movement through rectangular aperture, temp·2. The bars are oriented by $(45°)$ *and moved in upward direction* $(90°)$ *through rectangular aperture with aspect ratio 5:2. The V1 cell and one of the MT cells (V1:MT 1:3) are depicted over the accumulated events as blue circle and green circle, respectively. Motion representation in areas V1, MT and the modulated V1 for each stimulus are shown in the second, third and fourth column respectively. In this stimulus, the real motion direction* $(90°)$ *is estimated at bar endings which are distributed along the vertical edges of the rectangular aperture while the normal flow* $(135°)$ *is estimated along bar contour. The polar plot shows the direction selectivity of V1, MT and the modulated V1 responses which are depicted by blue, green and red lines, respectively. The histogram shows that the angular error between the estimated motion and the ground truth motion* $(90°)$ *is reduced in area MT and the modulated V1 compared with the initial motion estimation of V1 area (as shown in the fourth bin of error values* $[45°, 60°)$*). The abscissa of the histogram represents the binning in the range of the angular error* Φ *which are combined into one bar* $[\theta - 7.5°, \theta + 7.5°)$*, and the ordinate represents the number of events.*

erated at the corners regions. As a consequence, the direction selectivity of MT cells shrinks toward the real movement as shown in the polar diagram of the direction selectivity. According to the angular error histograms, values of the angular errors between the synthetic ground truth of the object movement and the estimated motion are decreased in area MT comparing with area V1.

In the case of rotational motion, we used smooth-cross and plaid stimuli, Figure 3.12. These stimuli were rotated counter clockwise, as highlighted in the top-left of the stimulus images. The smooth-cross stimulus contains four blades with smoothed gray-level interior. These blades produce changes in intensity during the rotational motion, rendering the DVS sensor to generate ON/OFF events on the edges as well as the interior of the blades. The plaid stimulus is shaped as two sets of black perpendicular bars on white background. The generated motion components represent a challenge in terms of calculating motion direction. This is due to each set of bars generates different stream of rotational component that suffers locally from aperture problem. The input stimulus and the respective ON/OFF events are presented in the first and second rows of the first column respectively. In order to calculate the angular error for the feedforward estimated motion, we used synthetic rotational ground truth that was built based on continuous flow motion as shown in the third row of the first column. The results reveal that motion prediction via V1 area is improved in area MT in which the magnitude of the error-bins ($\Phi \geq 45°$) and ($\Phi \geq 30°$) are decreased in the smooth-cross and plaid stimuli, respectively. In the rotational sweep, the high temporal of input events delivered by the DVS sensor leads to motion components that can be considered as to mainly represent motion components tangential to a rotational sweep. However, Cells in area V1 can only measure the normal flow motion components (movement orthogonal to the bar contrast). Cells in area MT with larger receptive fields integrate these components resulting motion representation that are tangential to rotational sweep. The direction selectivity of the model areas V1 and MT are presented in the polar diagrams as shown in Figure 3.12. The polar diagram shows the direction selectivity in area MT is highly tuned toward the correct direction. According to the error histogram, the error values are reduced in area MT and accumulated in the error-bins of [0, 30). The reason for the high error value in rotational motion is that the rotational ground truth was built based on continuous flow motion, while our model estimates eight directions. Thus the error value can be decreased by increasing the number of estimated directions in our model.

Since the rotational motion contains different speeds as a function of the radius from the center of the motion, we utilized a rotational bar stimulus to demonstrate the speed selectivity of the MT cells, see Figures 3.13 and

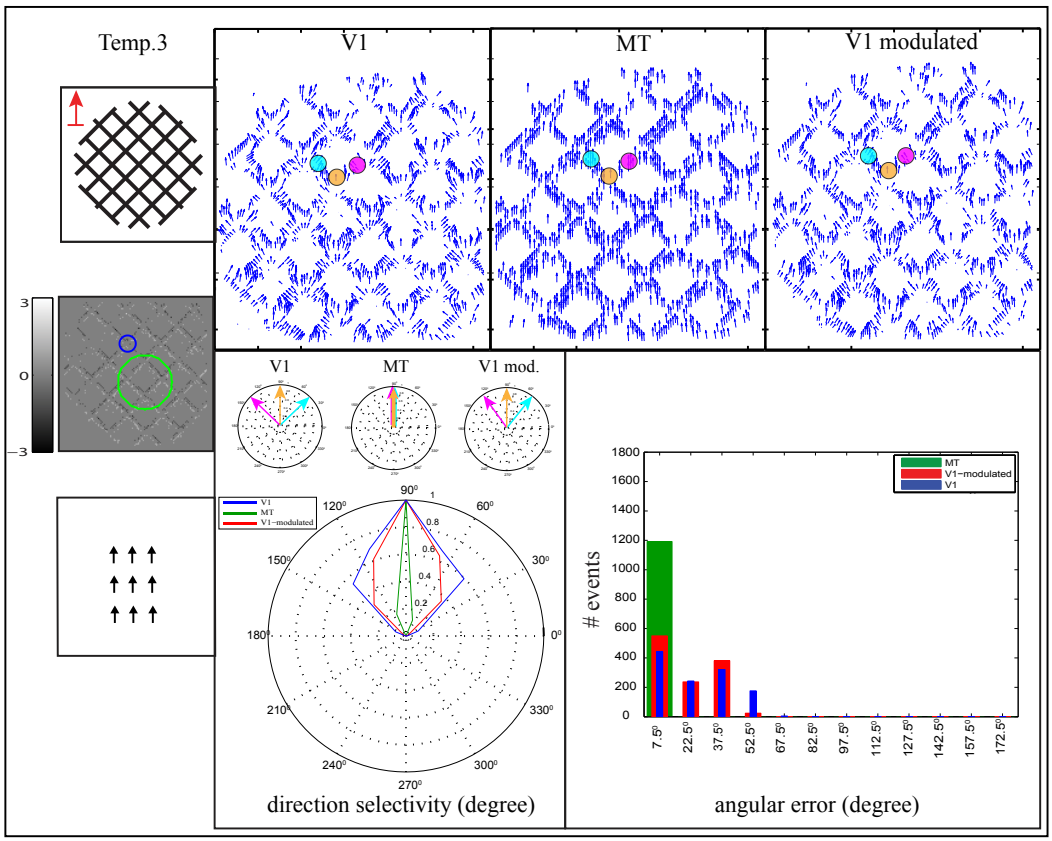


Figure 3.11: Motion estimation of 2D structure of corner terminators, temp·2. Two superimposed gratings are moved upward direction through circular aperture. The image input is shown in the first column of the first row. V1 cell and one of the MT cells (V1:MT 1:3) are depicted over the accumulated events as blue circle and green circle, respectively. The upper row shows the estimated motion in areas V1, MT and the modulated V1. The real motion is estimated at the 2D features (corner regions) while the normal flow is estimated along bars contours. The small polar plots show the direction selectivity of selective cells that are located on the bars contours and corner while the whole direction selectivity of the stimulus are depicted in the large polar plot. The histogram shows the angular error between the estimated motion and upward motion ground truth, where the abscissa represents the binning in the range of the angular error Φ which are combined into one bar $[\theta - 7.5°, \theta + 7.5°)$, and the ordinate represents the number of events.

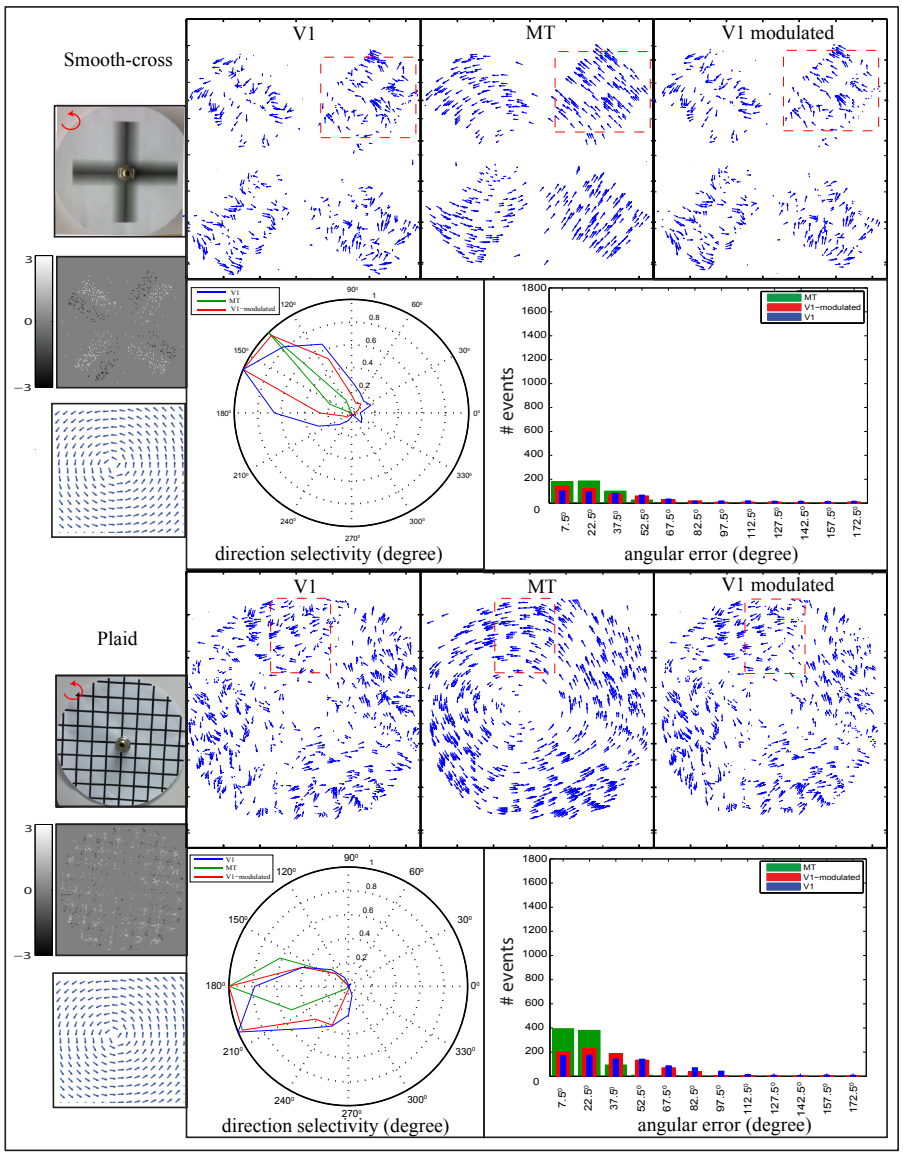


Figure 3.12: Motion estimation for smooth-cross and plaid stimuli. The stimuli are rotated in a counterclockwise direction. The first column of each stimulus contains the input image, accumulated events e^{on} and e^{off}, and the ground truth optical flow field. The first rows of the second, third and fourth columns represent the estimated motion in areas V1, MT and the modulated V1, respectively. The polar plot shows the direction selectivity of V1, MT and the modulated V1 of the bounded region (red square). The overall errors between the estimated motion and their respective ground truth are depicted in the histograms where the abscissa represents the binning in the range of the angular error Φ which are combined into one bar $[\theta - 7.5°, \theta + 7.5°)$, and the ordinate represents the number of events.

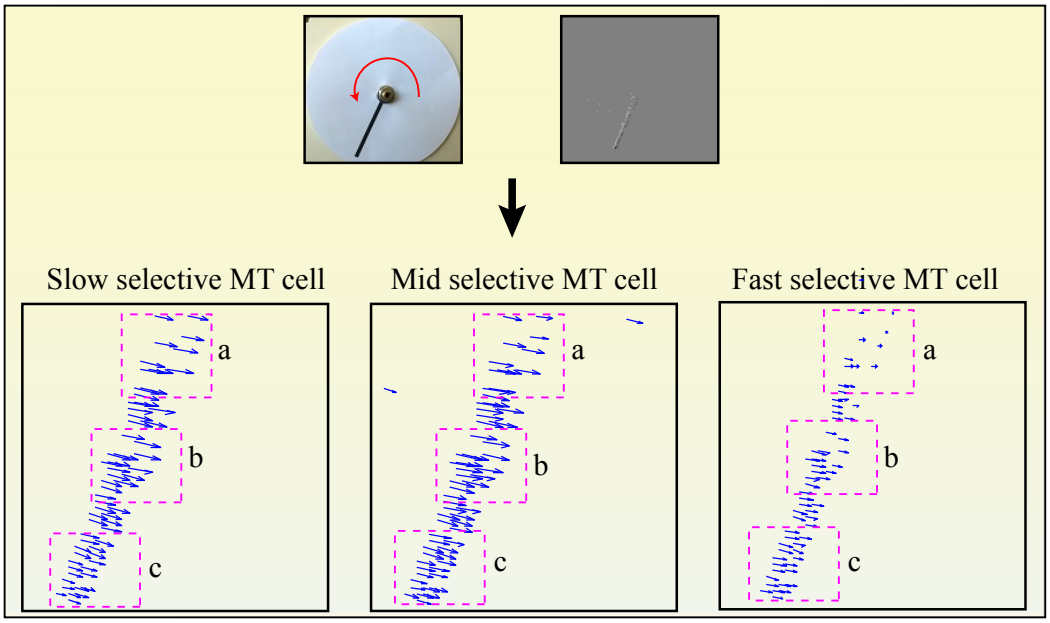


Figure 3.13: Speed selectivity of MT cells. A black bar (5 × 80 mm) is rotated counterclockwise on a white background. The first row shows the input stimulus (left) and the ON/OFF events (right). The second row shows the responses of the slow, mid and fast selective cells respectively. Here, the average speeds of thee regions (a, b and c) were calculated. In the slow selective cell, the average speed at region (a) is 0.7 while the average speeds at (b) and (c) are 0.6 and 0.4, respectively. In the mid selective cell, the average speed at region (a) is 0.6 while the average speeds at (b) and (c) are 0.7 and 0.5, respectively. In the fast selective cells, the average speed at region (a) is 0.3 while the average speeds at (b) and (c) are 0.5 and 0.6, respectively.

3.14. Here, three regions are chosen at the bar contour in which the first region is located closed to the center (a) while the rest are located away from the center (b and c). The speeds of these points are calculated based on the length of the flow vectors $\sqrt{(u_e(x,y,s_i))^2 + (v_e(x,y,s_i))^2}$ at the spatial position (x,y) and speed channel s_i. In order to keep a single motion hypothesis, the estimated results of the optic flow are generated based on a weighted sum of the fundamental directions θ, which generates a confidence for the motion direction in each speed channel s_i which can be described by $(u_e(x,y,s_i)\ v_e(x,y,s_i))^T = \sum_\theta r^{MTnor}_{x,y,\theta,s_i} \cdot (cos\,\theta\ ,\ -sin\,\theta)^T$. The results reveal that the slow sensitive MT cell is more selective to the speed of region (a) while mid and fast sensitive cells are selective to the speed at regions (b) and (c), respectively. These results confirm that cells in area MT can differentiate different speeds of the rotational motion.

3.5.2.2 Feedback motion representation

In the feedback sweep, cells in area V1 receive feedback signals from the higher area MT such that the feedforward activities can be modified via the modulatory influence of such top-down signals. Here, the feedback signals alone cannot excite any activation when feedforward signals are absent. In order to evaluate the impact of the MT feedback signals to enhance the initial estimation of area V1, we used the former stimuli that are used in feedforward motion representation in which the results of the modulated V1 are illustrated after 12 feedback iterations. In case of translatory motion, Figure 3.8, the motion estimation of the modulated V1 is shown in the fourth column. The modulated motion of area V1 in tiger stimulus shows that the re-entered signals form area MT to V1 cells enhance the initial estimation of area V1 in which the magnitude of error-bins [30°, 60°) is decreased along with increasing the magnitude of the smaller error-bins as shown in the error-histogram. In the ball stimulus, MT projection improves the initial prediction of area V1 cells in which motion ambiguity is reduced and amended to the right direction. In addition, the magnitude of the angular error-bins [30°, 60°) is decreased in the modulated responses while the magnitude of the first error-bin is increased. This confirm that the top-down feedback projection can enhance the initial motion representation of area V1 in which uncertain estimated flow is corrected.

To probe the impact of the MT feedback to reduce the ambiguity (aperture problem) of the initial estimation of area V1, we used the previous stimuli temp·1, temp·2 and temp·3. In the circular and rectangular apertures, Figures 3.9 and 3.10, the motion in area V1 is initially estimated orthogonal to the bar contrast (135°). This ambiguity is reduced in area MT because cells

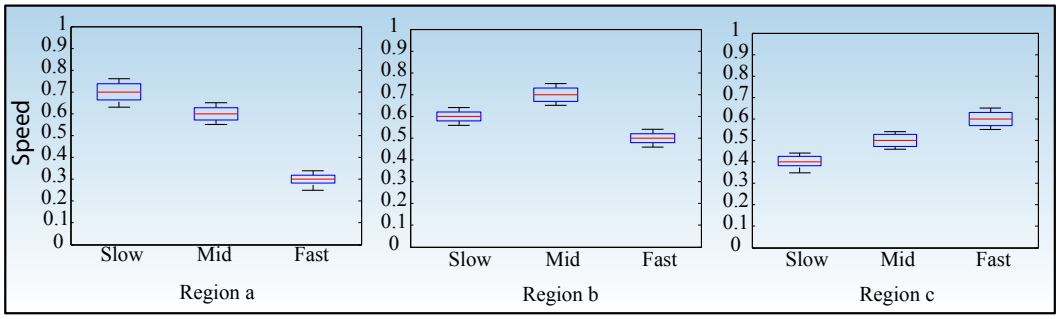


Figure 3.14: Response distribution of the speed selectivity MT cells. The box plots display the distribution of the responses that are generated by slow, mid and fast MT cells along three regions of the rotated bar shown in Figure 3.13. (Left image) response distribution of slow, mid and fast selective cells in region a. (Middle image) response distribution of slow, mid and fast selective cells in region b. (Right image) response distribution of slow, mid and fast selective cells in region c.

in this area integrate the V1 activations with larger RFs. As a consequence, the feedback signal which contains disambiguate contextual information modulates the local initial responses of V1 cells. The error histograms demonstrate that the motion ambiguity in the modulated responses of area V1 is enhanced in which the magnitude of the error-bins in the range of [45,60) is decreased. On the other hand, the existence of the 2D structure of the corners in the temp·3, Figure 3.11, reduces the ambiguity of the perceived motion in which the preferred direction of cells in area MT is more tuned toward the real motion direction. Thus, the top-down signal from area MT enhances the initial local estimation of area V1 where the direction selectivity shrink toward the actual motion (90°). The error-histograms show an increasing in unambiguous motion estimation where the error of [45°, 60°) is reduced and cumulated within a small range of [0°, 15°).

In the rotational motion, Figure 3.12, the results of the smooth-cross and plaid stimuli show that the feedforward response from the V1 population of cells is improved by feedback modulator of area MT. The direction selectivity of the modulated V1 responses is more tuned toward the correct direction comparing with the initial estimation of area V1. The error-histogram demonstrates that the error of the modulated V1 is reduced comparing with synthetic rotational ground truth where the magnitude of the error-bins [0,30) gains more increment.

In order to demonstrate the reduction in the spread of the population response that is derived after the modulatory feedback, we used two former

stimuli, tiger (translatory motions) and smooth-cross (rotational motion). Figure 3.15 (a and b) shows the box plot of the response distribution for the tiger and smooth-cross stimuli, respectively. The motion directions in Figure 3.15 are represented in the range of $[-180°, 180°]$ in which the left box plot shows the distribution of the initial motion estimation of area V1 and the right box plot shows the modulated V1 responses. In our model, the spatiotemporal filters generate a confidence for the occurrence of motion direction. In order to keep a single motion hypothesis, we obtained a weighted sum of the fundamental directions θ utilizing the estimated confidence value which can be described by

$$\begin{pmatrix} u_{x,y}^{V1,V1mod} & v_{x,y}^{V1,V1mod} \end{pmatrix}^T = \sum_{\theta} r_{x,y,\theta}^{V1nor,V1mod} \cdot (cos\,\theta\,,\ -sin\,\theta)^T\,, \tag{3.13}$$

where $r_{x,y,\theta}^{V1nor,V1mod}$ represents the confidence of the V1 and modulated V1 responses for motion direction θ at spatial position (x,y), $\begin{pmatrix} u_{x,y}^{V1,V1mod} & v_{x,y}^{V1,V1mod} \end{pmatrix}^T$ represent the optic flow of area V1 and the modulated V1. The single motion hypothesis is finally calculated by $\vartheta_{x,y}^{V1,V1mod} = tan^{-1} v_{x,y}^{V1,V1mod} / u_{x,y}^{V1,V1mod}$. Here, the resulting direction ($\vartheta_{x,y}^{V1,V1mod}$) is used as a data base to generate the box plot. The results reveal that the spread of the modulated V1 responses is significantly reduced (paired t-test $P < 0.05$ and $P < 0.001$ in the tiger and smooth-cross, respectively) compared with the initial estimation of the V1. This confirms that the feedback projection can enhance the initial motion estimation of area V1 by reducing the spread of the response population and decreasing the angular error between the estimated motion and the ground truth. The results of our model are consistent with the findings of the literature (see e.g., [Desimone and Duncan, 1995]; [Bayerl and Neumann, 2004]; [Tsui et al., 2010] and [Bouecke et al., 2011]) in which the top-down modulatory signals emphasize the activities of V1 cells, which in turn, enhance the uncertain flow estimation.

3.6 Discussion

3.6.1 Summary and main contributions

We have introduced a new model for motion estimation in neural architecture utilizing the primary stages of the dorsal pathway in the primate visual cortex. We take into account event-based input that has been generated by a DVS sensor (see [Delbruck and Lichtsteiner, 2008] for more details of DVS sensors).

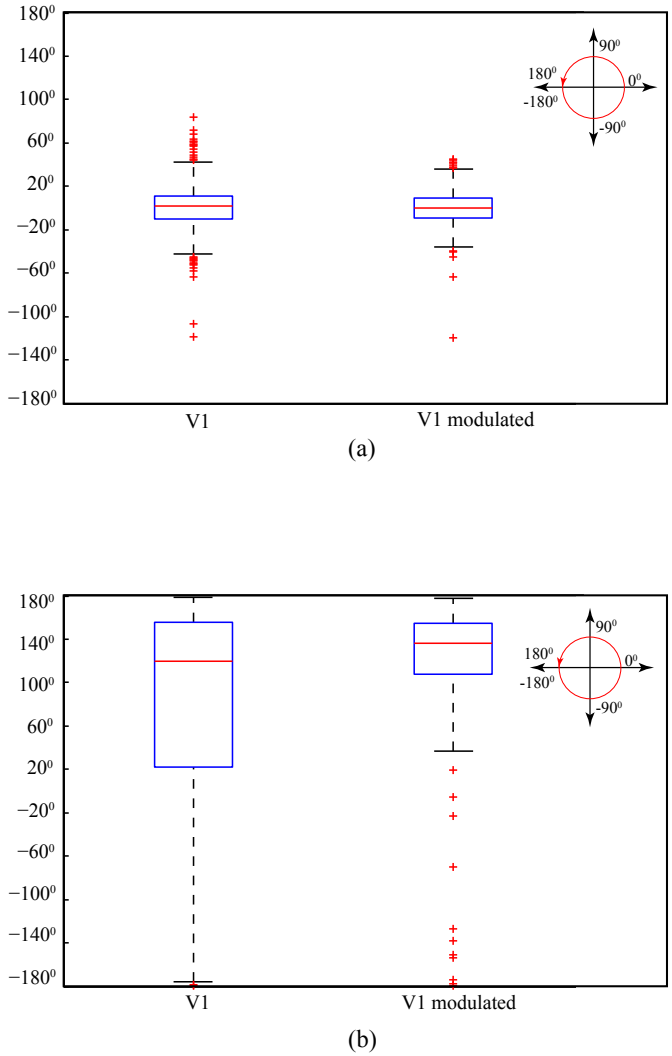


Figure 3.15: The distribution of area V1 responses and the modulated V1 responses. (a) The box plot of the response distribution in the tiger stimulus. (b) The box plot of the response distribution in the smooth-cross stimulus. In each image, the left box shows the distribution of the initial responses of area V1 while the right box plot shows the distribution of the modulated V1 responses. The motion directions are represented in the range of $[-180°, 180°]$ *as depicted in the top-right of each image. The distribution diagrams show that the spread of the V1 responses is significantly reduced (paired t-test* $P < 0.05$ *and* $P < 0.001$ *in the tiger and smooth-cross, respectively) after the modulatory feedback.*

The model is inspired by the hierarchical structure of the visual system in which two cortical areas for motion perception were considered, namely the response properties of V1 and MT. These areas interact via feedforward and feedback connections in which each area builds up different feature selectivities. Areas V1 and MT were modeled using properties of direction-selective and velocity-selective cells, respectively. Here, the spatio-temporal filters are modeled to be compatible with the AER principle. The filters are inspired by the work of [Adelson and Bergen, 1985] in which the authors proposed biphasic temporal filters that can be decomposed into a convolution of a first order derivative kernel with a temporal smoothing filter. Given that the response of the DVS sensor is based on changes in the luminance of the scene, it generates an output response which related to a first order temporal derivative of the luminance input signal.

This work contributes to the state of the art in four main ways. **First**, by adapting the spatio-temporal filters of the type suggested by findings of [Adelson and Bergen, 1985]. Here, we employ smoothing filters that are generated from the integral operation of the temporal gamma function (see Equation 2.10). To simplify the mathematical description of temporal filters in Equation (2.10), we suggested to combine two Gaussian functions to closely resemble the shape of the Adelson-Bergen temporal filters. **Second**, by proposing a new mechanism to simulate the functionality of the cortical area MT in which the responses of the area V1 are weighted and integrated in such a way to enable the model to detect different speeds as well as directions. This is accomplished by using MT cells with circular and elongated RFs. The elongated RFs are oriented in a space-time domain with different angles to encode the speed of the input stimulation in which high oriented angles represent high-speed detectors, while small oriented angles represent slow-speed detectors. **Third**, by mapping the properties of the feature selectivity in areas V1 and MT for feedback processing. Here, we integrated the responses over different speeds along a particular preferred direction. Such mapping allows top-down feedback signals from area MT to match feature property of area V1. **Fourth**, by incorporating the response normalization in our model to achieve balance activities of individual neurons in areas V1 and MT, in which this normalization is carried out in spatial as well as feature domains.

As pointed out below (Section Feedforward and feedback interaction) we incorporate mechanisms of modulating feedback which enhances driving feedforward activations. In a nutshell, the architecture enhances signals which match along the feedforward and the feedback pathways while those feedforward signals that do not cohere with top-down predictions are reduced. In order to augment the modulatory feedback mechanisms with a mechanism to

reduce and even extinguish activations, a context sensitive down-modulating mechanism needs to be employed. The proposed normalization stage that operates upon a pool of cells in the surrounding neighborhood of spatial and feature selectivity serves a mechanism to reduce the overall activation. The normalization tends to conserve the overall signal energy in the pool of cells which competitively interact. The prior enhancement of selected cells in turn reduces the activity of those cells that have not received any feedback.

3.6.2 Feedforward and feedback interaction

The feedforward and feedback hierarchical model in this work considers two main areas of the visual system, V1 and MT. The structural model of each of such visual areas is defined by three stages: (i) a stage of initial input filtering, (ii) a stage of activity modulation of the filtering responses via top-down feedback signals, and (iii) a stage of activity normalization in the spatial and feature domains to achieve balanced activations of a target cell against a pool of neighboring cells. The structural model proposed here can be transcribed to the cortical areas architecture that are suggested in [Self et al., 2012]. Each of such areas has a specific filtering stage model to generate the driving feedforward signal with particular features. Feedforward signals propagated from area V1 drives the direction feature while the feedforward signal of area MT drives the velocity feature (direction and speed). This framework of feature selectivity of areas V1 and MT is consistent with experimental findings [Lagae et al., 1993, Simoncelli and Heeger, 1998, Perrone and Thiele, 2001]. The activations of each area are normalized using a divisive mechanism suggested in [Carandini and Heeger, 1994]; [Carandini et al., 1999]; [Kouh and Poggio, 2008]. Here, the same principle is utilized in which the normalization operation uses contextual information from a local neighborhood that is defined in space as well as feature domain.

Higher-level areas in the visual cortex send feedback signals that are reentered to the earlier areas in the visual hierarchy [Grossberg, 1980]; [Edelman, 1993]. Different hypotheses for cortical feedback have been discussed in the literature in which two major hypotheses have received different support from the experimental evidence [Markov et al., 2014]. These hypotheses are defined as modulatory (biased competition) and predictive coding feedback. In a nutshell, modulatory feedback suggests that signals in the feedforward stream are enhanced by feedback projection. This feedback projection driven gain control mechanism to bias subsequent competition between neurons which leads to enhance responses patterns [Girard and Bullier, 1989]; [Desimone, 1998]; [Roelfsema et al., 2002]; [Reynolds and Heeger, 2009]. While predictive coding aims

to reduce the residual error between the feedforward signals and feedback projection in order to approach sensory prediction that generated via higher level of processing [Ullman, 1995]; [Rao and Ballard, 1999]; [Bastos et al., 2012]. Further evidence shows that feedback processing tends to act as a modulator that amplifies the neuronal spiking signals at the level of cortical pyramidal cells [Larkum, 2013]. In our model, we adopted the concept of modulatory feedback (biased competition). Here, the responses from lower level cortical area V1 are modulated via feedback projection of the higher area MT. Feedback modulation process is thereby carried out in such a way that feedback signals could not provoke any activity in the absence of the feedforward signals.

Although, the neuroscientists' studies confirm the impact of feedback processing stream among cortical visual areas, the precise function of feedback role is still not fully uncovered. A fundamental question here is how MT cell responses project activations along their feedback stream to area V1, since cells in area V1 are selective for direction while cells in area MT are selective for both speed and direction? This question directly addresses the suggestion made by [Edelman, 1993] that maps in different (cortical) area re-enter their signal. We proposed integrating the responses over different speeds along a particular preferred direction. As far as we know, this is the first event-based mechanism to address feedforward and feedback interactions between two areas that have different feature selectivity. In addition, response normalization for areas V1 and MT was embedded to obtain balanced activities in both areas. Here, we implemented the normalization at equilibrium state. The activities of the cells were normalized in the spatial domain through a distance weighting dependency filter (fall-off function), where the size of the filter is larger than the size of the cell RFs. In addition, we regulated the responses over feature space by calculating the average activity over all directions.

3.6.3 Relation to previous models of motion estimation

Motion estimation is an interesting topic that is investigated intensively using conventional frame-based cameras (see, e.g., [Adelson and Bergen, 1985]; [Lucas and Kanade, 1981]; [Horn and Schunck, 1981]; [Fleet and Jepson, 1990]; [Brox et al., 2004]; [Drulea and Nedevschi, 2013]). Relatively few studies have been reported how to transcribe the functionality of such classical approaches to be consistent with neuromorphic vision sensors (see [Lichtsteiner et al., 2008] for more details about neuromorphic vision sensors). In [Benosman et al., 2012] the authors used a least squares error minimization technique introduced in [Lucas and Kanade, 1981] to estimate the motion using an DVS sensor. Due to DVS sensor generates a stream of events (ON or OFF) and

does not provide gray levels, thus, the authors have been suggested to use pixel activities by integrating events within a short temporal window. Benosman and co-authors showed beneficial results for motion estimation, however, their numerical approximation of the local gradients of the luminance function from event-sequences has its limitations and may lead to inconclusive results (see [Brosch et al., 2015]; [Tschechne et al., 2014a]).

In [Tschechne et al., 2014b], the authors introduced an algorithm for motion estimation using a DVS sensor in which spatiotemporal filters of the type suggested by findings of [De Valois et al., 2000] were utilized to estimate a local motion for each generating event in the scene. The authors implemented the spatiotemporal filters using a spatial buffer in which the timestamp of each event is stored. Recently, Tschechne and co-authors extend their work in [Tschechne et al., 2014a] by introducing a framework of a hierarchical architecture of multi-stage motion detection and integration in which V1 responses are integrated in area MT over a larger neighborhood using circular RFs. Here, the authors have focused on motion direction selectivity in area MT. The implications of event-based sensing in the context of visual motion have been investigated by [Brosch et al., 2015]. Brosch and co-authors discussed different principal approaches for optical flow detection. They showed that gradient-based methods for local motion detection in principle suffer from the sparse encoding in address-event representations (AER) because they are rare with respect to a local weighted integration during filtering. The authors further investigated approaches to exploit the local plane-like structure of the event cloud and how local filtering can be properly defined.

The weighted intersection mechanism (WIM) sensor has been proposed by [Perrone and Krauzlis, 2002] and developed in [Perrone, 2004]. Here, a motion sensor built up in stages from two spatiotemporal filters with properties based on V1 neurons. The sensor mechanism incorporates two V1-like units based on spatio-temporal energy filters. The first unit has sustained low-pass temporal frequency tuning (referred to non-directional type) , whereas the second unit has transient band-pass temporal tuning (referred to directional type). This mechanism enables two filters with broad temporal tuning (one low-pass and the other band-pass) to be converted into a filter with tight temporal frequency tuning and an orientation that maps onto the oriented spectra generated by moving edges. Perrone and co-authors showed that the speed tuning property of such a WIM filter is comparable to that found in many MT neurons.

Feedback processing tends to act as a modulator input from higher areas that mediate top-down contextual effects [Friston and Büchel, 2000]. Such processing has been demonstrated to enhance the gain of neural representa-

tions during different processing phase in several tasks such as texture segregation (figure-ground segregation) [Poort et al., 2012] as well as increasing the visual awareness [Mumford, 1991]. The feedback modulation hypothesis, No-Strong-Loops, implements a distinct driving and modulating inputs principle (see [Crick and Koch, 1998]). The driving input can strongly activate the neurons that are concerned with sensory input processing, or filtering. The modulatory inputs, on the other hand, cannot generate any activity by themselves but can modify the driving input of a target neuron utilizing contextual information provided by higher-level representation. The modulatory feedback is considered as a common principle that has been observed in many studies. In [Bayerl and Neumann, 2004]; [Bouecke et al., 2011] a model of motion processing in areas V1 and MT has been proposed in which driving feedforward and modulating feedback signals interaction are considered for the purpose of motion detection and integration. In the model, the localized motion response in area V1 was integrated in area MT via feedforward processing stream while in feedback processing the initial estimation of area V1 was modified via feedback projection from area MT. The authors showed that top-down modulatory signals emphasize the activities of V1 cells, which in turn, enhance the uncertain flow estimation and improve the visual motion segregation. In [Layher et al., 2014b], the authors presented a hierarchical architecture of cortical feedforward and feedback computation where they proposed how a top-down feedback-modulatory learning mechanism can increase the gain of the feedforward driving inputs. This results in network feature representations that automatically adjust the connection weights utilizing unsupervised learning. As a result the top-down predictions are improved for the input pattern.

The motion boundary contour system (or Motion BCS) model have been introduced in [Chey et al., 1997] and [Chey et al., 1998]. The authors suggested a neural model for the purpose of motion detection and integration. The input of the model taken to be the outputs of FACADE mechanisms (Form And Color and DEpth processing) which is firstly described by [Grossberg, 1994]. This model has been extended by [Grossberg et al., 2001] where several processing stages with a feedback interaction between MT and MST stages are presented. Grossberg and co-authors showed how the aperture problem can be solved based on 2D feature signals. This has been accomplished by computed unambiguous motion from feature tracking points which are amplified before they propagate across position and are integrated with ambiguous motion signals within bar interiors. Grossberg and co-authors demonstrated how the feedback signals from model MST to MT cells which encode the winning direction boost directionally consistent cell activities and suppress inconsistent activities over the spatial region to which they project. A model of recurrent

motion processing between areas V1 and MT has been suggested by [Bayerl and Neumann, 2004], where the authors suggested a model of V1-MT feedforward and feedback processing in which the model of each area consists of three steps, namely feedback modulation, feedforward integration, and lateral inhibition. The initial motion detectors in area V1 consists of a set modified elaborated Reichardt detector. In feedforward processing the localized motion representation in area V1 were integrated via model area MT using larger receptive fields (V1:MT, 1:5). In the feedback processing the initial estimation of area V1 were modified via the feedback projection from area MT. The authors showed how ambiguities of detected visual motion can be solved by combining mechanisms of local lateral interaction with modulatory feedback.

In [Bouecke et al., 2011], a model of motion processing in areas V1 and MT has been proposed in which driving feedforward and modulating feedback signals interaction are considered. Here, the localized motion response in area V1 were integrated in area MT via feedforward processing signals while in the feedback processing signals the initial estimation of area V1 were modified via the feedback projection from area MT. The authors showed that top-down modulatory signals emphasize the activities of V1 cells, which in turn, enhance the uncertain flow estimation and improve the visual motion segregation.

Our model differs from these other approaches in the initial method of motion detection. Here, we adopted the bio-inspired model suggested in [Adelson and Bergen, 1985] and adapted the filtering principle to make the approach consistent with the functionality of DVS sensors. The event-based technology of visual sensors provides our model with a high temporal resolution (1 μs). In addition, the redundant information captured by conventional frame-based camera is reduced. The architecture of the proposed model takes into accounts feedforward and feedback interactions in which area V1 are modeled as direction selective cells while area MT modeled as velocity (speed and direction) selective cells. Here, V1 responses are integrated via elongated RFs that are tilted in space-time domain. This kind of integration increases the cells sensitivity to detect different speeds, fast, mid and slow. Our model take into accounts feedback processing and focuses on how context information from higher-level area can re-enter into lower-level area in which both areas have different feature selectivity. This has been done by integrating the responses of MT cells over different speeds to match the properties of both areas, V1 and MT.

3.6.4 Model evaluation and future work

To verify the speed sensitivity of the proposed elongated RFs, a translating bar with different speeds (slow 7 cm/sec, medium 17 cm/sec and fast 23 cm/sec) was used. The results demonstrate the correct speed sensitivity of the suggested cells. Our model was tested using different kinds of stimuli moving in different directions. In feedforward processing, the results show that V1 population responses are selective to broad directions. These responses are enhanced in area MT in which the responses are highly tuned to the correct directions. This outcome is consistent with the functionality of MT cells which increases the direction selectivity (see e.g., [Churchland et al., 2005]). Feedback modulation processing, on the other hand, improves the direction selectivity of the initial motion estimation via area V1. However, the aperture problem has not been completely solved. Further work is needed to resolve this issue.

We demonstrated the angular error for the bottom-up and top-down predictions using a histogram, where the error was calculated by comparing the estimated flow with synthetic ground truth. The stimuli results show an improvement in the flow estimation in area MT and modulated V1, compared with the initial motion estimation in area V1.

Our model can be extended by adding other functions of the cortical areas such as medial superior temporal (MST). This will bolster the models ability to process more complex motions in which the cells in this area have larger RFs than area MT and can respond to complex patterns of visual motions. In addition, the projection feedback from this area to MT will enhance the response activities of MT cells. On the other hand, the model can be developed by carrying out the normalization process in the temporal domain. This can be accomplished by pooling the activities of the V1 and MT in the temporal domain using Gaussian weighting function. The normalization process in the temporal domain balance the activations of the individual cells against the pool of activation over time axis.

The architecture of our model can be used as a basic adaptive scheme for motion estimation based on sparse event-based input. It is thus conceivable that other researchers interested in biologically inspired technology based on address event representation may start from this point in order to further develop mechanisms in this framework. In Chapter 4, we introduce the usability of the neural mechanisms defined here for classification purposes. We considered articulated motions for this purpose in which biological motions are selected as a case study.

Chapter 4

Articulated motion recognition

4.1 Introduction

Despite several decades of research, articulated motion recognition is still a challenging task in computer vision. Such motion can be generated from the human body, many enamels and the movement of machinery. Motion recognition is closely related to several application domains such as human action recognition, surveillance and object tracking and obstacle detection. Here, we consider human action recognition by focusing on the movement of human limbs and the body.

Establishing an automatic visual recognition system for biological motion requires many assumptions and constraints. Various factors can affect the action recognition. In human action recognition, different features can be generated when the same action is performed from different view points. In addition anthropometry, human complexion and clothes have an impact on appearance and consequently on the action recognition. The camera sensing specifications are also relevant, such as sampling rate and image resolution.

Conventional frame-based cameras have been used for many years for biological motion acquisition. This kind of camera captures action scenes as a series of full frames which most likely contain non-contextual information of object motion. As a consequence, these frames require substantial preprocessing in order to extract the motion features of an action. In contrast, event-based cameras generate spikes in response to luminance changes in the scene, where the DVS sensors simulate the transmission of biological retinal responses. In this chapter, we investigate how the feedforward and feedback model, defined in Chapters 2 and 3, can be utilized as a platform for human action recognition tasks using event-based sensing.

.

4.2 Human action recognition

Human action recognition can be defined as a classification process that deals with understanding human activity. This activity is recognized through the analysis of temporal visual data. Over time, our understanding of human action-perception has advanced significantly, where the perception of an action depends on multiple sources of information including sensory, motor, and affective processes [Blake and Shiffrar, 2007].

Action recognition has been addressed in neuroscience and psychophysics showing that human activity does not necessarily need the moving scene of a full actor. A notable study of [Johansson, 1973] showed that human activity could be recognized through point-light stimuli which correspond to human joint positions. Point-light stimuli of biological motion have been used in many psychophysics studies to gain a better understanding of the underlying mechanism of the perception of human action recognition (such studies are addressed in Section 4.2.1). However, neural mechanisms of biological motion recognition have in fact been discussed extensively in the literature. The contribution of the *form pathway* or *motion pathway* in the perception of human action recognition remains unclear. Some studies (see for e.g., [Beintema and Lappe, 2002]) have shown that biological motion can be defined using *form* information of the human body. On the other hand, [Casile and Giese, 2005] showed that motion information has a significant impact on the perception of biological motions. Other evidence (e.g., [Hirai and Hiraki, 2006]) has confirmed that the motion pathway of the visual system plays a key role in biological motion recognition, though other brain areas of the form pathway and feedback processing are also involved in the perception of human action. Neuroscientists have investigated the neural mechanism of the brain architecture via fMRI (see [Michels et al., 2005]) and confirmed that both pathways, form and motion, are involved in the analysis of the biological motion.

Bio-inspired models for biological motion recognition are still a challenge in neurosciences. Here, the main question is how can action scenarios be transformed into essential traits for action recognition based on a bio-inspired vision model? The aim of our work is to describe biological motion via informative features that can be used to identify actions. These features rely on the motion energy of the body and limbs.

4.2.1 Previous work

Motion processing is a significant feature in human action recognition and has been widely investigated in neuroscience studies. A considerable amount of literature has been published on the visual motion processing. The ar-

ticles related to our work have been reviewed in this section. The studies by [Johansson, 1973] and [Juhansson, 1976] have contributed significantly to biological motion discrimination based on visual system findings. The authors found that human visual perception can recognize biological motion through moving point-light stimuli placed at the joints of the human body. Several attempts have been made to extend the representation of the point-light stimuli to identify individual people or to distinguish gender (see, e.g., [Cutting and kozolowski, 1977], [Torje, 2002] and [Pinto and Shiffar, 1999]).

Various mechanisms have been developed to identify the characteristics of biological motion. In [Bobick and Davis, 2001], the authors considered two components of temporal templates, motion-energy images (MEIs) and motion-history images (MHIs), to describe where and how motion takes place. In [Jhuang et al., 2007], the authors presented a hierarchical feedforward model based on invariant space-time feature detectors, using the max-operator. [Yilmaz and Shah, 2008] used the spatio-temporal space to describe an action as a 3D volume through stacking a sequence of tracked 2D object contours. Object motion representation based on local-space-time descriptors is considered in [Laptev, 2005]. Laptev described interest points in spatio-temporal scale-space, where these points are characterized by a high variation in image values in space and non-constant motion over time. This study was extended in [Laptev et al., 2007] where the authors investigated event detection under Galilean transformations arising from the relative motions of the camera.

Several models have been developed to simulate the functionality of human visual system, where some authors focus on form analysis in addition to motion analysis such as [Beintema and Lappe, 2002]; [Niebles and Fei-Fei, 2007]and [Laptev et al., 2007]. The authors showed the prominence of the form features needed to improve biological motion recognition. In keeping with the importance of the information regarding the form pathway, [Schindler et al., 2008] constructed motion recognition of the language of the human body based on form-perception by considering body poses. In addition, [Casile and Giese, 2003], [Casile and Giese, 2005] and [Giese and Poggio, 2003] demonstrated the dominance of motion information in biological movement recognition and demonstrated that the form pathway is not excited by point-light-walker stimuli. Interestingly, neurophysiological studies through functional MRI have shown the biological motion excited both the form and motion pathways of the brain (see, e.g. [Vaina et al., 2001], [Michels et al., 2005]. Recently, [Escobar and Kornprobst, 2012] presented a biological motion recognition model via the motion pathway. The authors demonstrated how the MT center-surround interaction is a major advantage for action recognition. [Layher et al., 2014a] introduced a neural model for biological motion recognition consisting of two

separate visual pathways for segregated form and motion processing as inspired by [Giese and Poggio, 2003]. Layher and co-authors extended the model by combining it with hierarchical feedforward and feedback mechanisms along motion and form pathways.

Our approach contributes to bio-inspired models for action recognition using event-based sensing by extracting features of an action through spatial regions of interest over human limbs and the body. These features are derived from the motion pathway in which two main areas (V1 and MT) for motion estimation are used.

4.3 Our model for action recognition

This section describes the architecture of the biological motion recognition of our model. The biological motion is estimated using the neuronal mechanism of the feedforward and feedback scheme which is defined in Chapters 2 and 3. The architecture of the motion estimation model consists of three processing stages that roughly correspond to model areas of the visual cortex. The first stage represents the visual input data in which event-based sensing is used for data acquisition. The second stage defines the model mechanism of area V1 in which spatiotemporal filters are used to locally detect the direction of motion. In the third stage, the responses of area V1 are weighted and pooled by cells of the model area MT which are selective to different velocities. The second and third stages of the model hierarchy are connected bidirectionally representing the feedforward and feedback stream of signal propagation in the visual cortex. We have already explained in Chapter 3 how these two stages can be communicated even though different features derive from both stages. In this section, the output of our model in response to a biological motion are interpreted into traits that are later used as a feature map for action classification. The following subsections introduce the layout of the classification mechanism.

4.3.1 Feature vector

In the previous section, we briefly described the main stages of our proposed model which is inspired by the findings of neuroscientists. The question thus arises: how to define essential features that can identify a human activity? We suggest using the output of the third stage (area MT) to extract a feature vector of human movements. The contextual information of the feature vector is derived based on events cloud of the input stimulus. These features are then fed into a supervised classifier, as presented in Section 4.3.3 for action

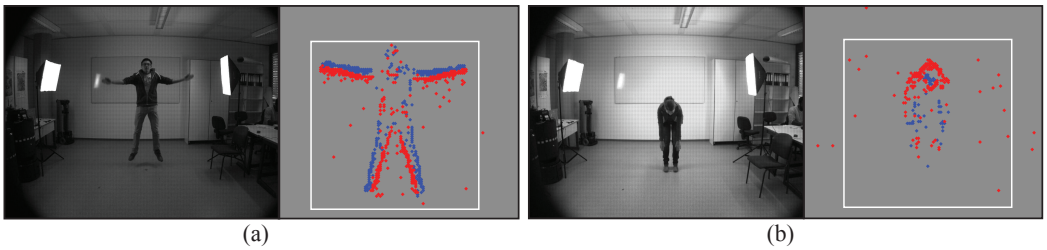


Figure 4.1: Focus on human movement. (a) jumping-jack action. (b) bending action. The white frame represents the boundary window of human activity. The red points represent the ON events while the blue points represent the OFF events

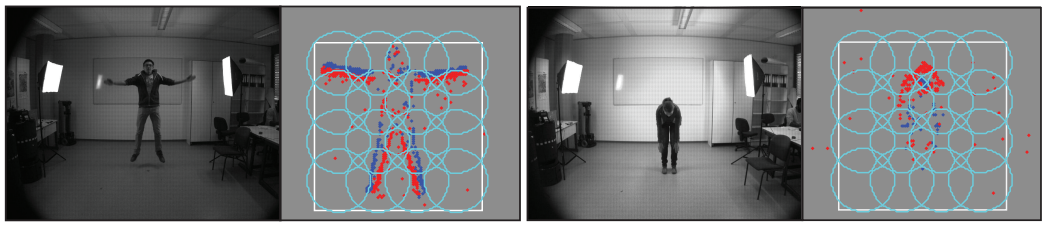
recognition. Our mechanism for feature extraction roughly follows the concept suggested by [Escobar and Kornprobst, 2012]. In the following steps, we introduce the description of our mechanism for feature extraction:

- ***Focus on action***

  In visual perception, follow-up human movement and focus on actions are necessary for action recognition (see [Orban de Xivry et al., 2010]). To focus on biological motion, an attentional mechanism is required to select part of the visual scene and hence handle a specific portion of the sensory information. We use a boundary window, which is located in the center of the human movement, to focus on an action in the visual scene. The bounded area of the visual scene is analyzed to extract the features vectors of the biological motion input. The 2D size of the boundary window is defined by (M $\times$ N). Figure 4.1 shows two examples of the boundary windows located on the sparse events representation of jumping-jack and bending actions.

- ***Motion dynamics via a spatial sampling detector***

  The motion dynamics of the human body and limbs can help in estimating motion behavior and action. In [Datta et al., 2011], the authors estimated the body motion based on an articulated constraint plain. On the other hand, [Bauckhage et al., 2009] estimated normal and abnormal gait patterns via the shape of the human motion using a bounding box splitting lattice. [Gorelick and Blank, 2007] demonstrated the action dynamics based on space-time shapes in the space-time volume which contain spatial information (the posture of a human figure at any time), as well as dynamic information (global body motion and motion of the limbs relative to the body).



(a) (b)

Figure 4.2: Motion dynamics detection of human activity. (a) jumping-jack action. (b) bending action. The white frame represents the boundary window of human activity. The light-blue circles represent the sampling regions that are located over the visual view of human movement.

In our work, the motion dynamics of an action are identified through a number of spatial sampling regions. We sampled the spatial motion of the human body and limbs via a $(M_c \times N_c)$ grid of coarse sample parts with an offset. Here, the spatial sampling regions are designed as circular masks and placed on the window. Figure 4.2 shows two examples of the ON/OFF events for jumping-jack and bending actions highlighting the organization of masks over the human body and limbs. The biological motion is described through the variation magnitude of the motion within each mask. The upper and lower masks can detect the motion of limbs (arms and legs) while the overall motion of the body can be detected by the median columns of masks. The overlapping areas of the masks maintain the sampling theorem and no spatial information is lost.

In our mechanism, the feature vector is defined by H^{av} whose components are the average activity within each area (A_c) of the sampled regions. The average activity is calculated for each motion direction (θ). The feature vector H^{av} thus can be defined by

$$\forall r^{MT}_{\theta,s_i,x,y} \in A_{c_k}; H^{av}_{\theta,s_i,c_k} = \frac{1}{p_n} \sum_{j=1}^{p_n} r^{MT}_{j,\theta,s_i,x,y}, \tag{4.1}$$

where $r^{MT}_{\theta,s_i,x,y}$ denotes the responses of area MT which are derived from different directions θ and speeds $(s_i = \{slow, mid, fast\})$, p_n defines the number of the activated cells within a specific sampled region (c_k) in which $k = \{1 \dots M_c \times N_c\}$.

4.3.2 Experimental protocol

In order to evaluate the performance of our mechanism, we used a data-set recorded by our vision group at the Institute of Neural Information Processing-Ulm University. This data set contains several biological motions, or actions where 11 different samples of different human subjects performing seven actions are selected. These actions are: walking on the spot (*walk*), running on the spot (*run*), waving one hand (*wave1*), picking up an object from the ground (*pickup*), bending (*bend*), upper body pushing exercise (*pushups*), and jumping-jack (*jack*). These actions were recorded using the neuromorphic vision sensor (DVS128-iniLabs). The vision sensor was mounted on a tripod and centered on the motion view, Figure 4.3 shows the sparse events representation of the seven selected biological motions.

These actions were selected to represent different human activities that permeate our daily lives. The selected actions contain different motion velocities (directions and speeds) of the human body and limbs. The actions were recorded in indoor environments using an artificial ambient light. Both males and females perform these types of actions and scenarios. The participants wore different apparel with different colors. This enabled us to evaluate the robustness of our model for motion detection and classification. We applied our classification mechanism for all these actions some of whose motion information was used for training, while the rest was kept out for testing. In the following section, we introduce the details of our classification mechanism.

4.3.3 Supervised learning of classifier

In the previous section, we introduced a description of the model mechanisms for feature extraction. As already mentioned, the feature vectors are extracted through spatial sampling of the human body and limb motions. The spatial sampling are located on a bounded area of the visual scene. We considered the common strategy of supervised classification in which some of the inputs were used as a training class (also known as a training set). On the other hand, the remaining inputs were used for testing operations (known as the test set) where an input of the test set was analyzed to extract the corresponding feature vector which was then classified to the closed class.

Support vector machines (SVMs) are one of the most widely used classifiers for object recognition (see e.g. [Linde and Lindeberg, 2004, Nilsback and Caputo, 2004]; [Wallraven et al., 2003]; [Wolf and Shashua, 2003]) and are also an effective classifier for human actions [Laptev et al., 2007]; [Jhuang et al., 2007]; [Yeffet and Wolf, 2009]; [Wang et al., 2010]. In our work, an *all pairs SVM* strategy was used for action classification. This strategy (also known

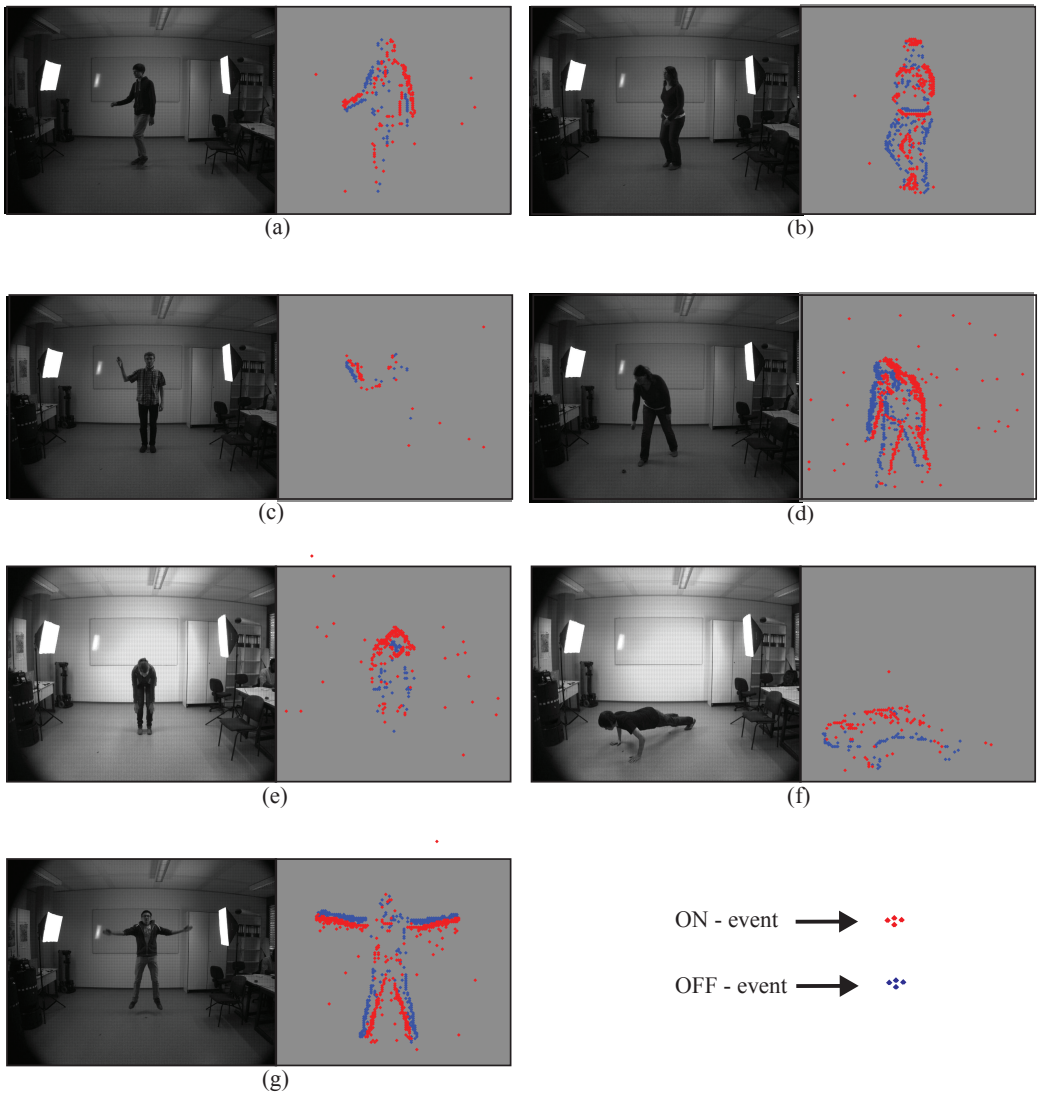


Figure 4.3: biological motions. (a) Walking in spot. (b) Running in spot. (c) Waving one hand. (d) Picking up an object from the ground. (e) Bending. (f) upper body pushing exercise. (g) jumping-jack. The left image of each action represents the frame based image while the right side shows the ON/OFF events in response to the human motions.

as *pairwise coupling*, *one against one*, or *round robin*) consists in constructing one SVM for each pair of classes [Milgram et al., 2006]. Hence, a classification operation for n classes requires to train (n(n-1)/2) SVM in order to recognize the patterns of one class from the other class patterns. Each SVM votes for one class, and as a consequence, recognition of unknown samples is achieved according to the maximum voting class.

We used a leave-one-out-procedure to evaluate the performance of the classification algorithm, in which one sample is excluded from the data set and the remaining samples are used to drive the recognition functions. This is systematically repeated such that each sample is used for testing once. In our evaluation process, the recognition performance was estimated over all possible combinations of training sets, i.e., 11 trials. As each subject has particular features, we removed the test subject from the whole actions by which no indication of subject features in the training sets can exist, which leads to a reliable conclusion about the performance. In the following section we introduce the results of our classification algorithm.

4.3.4 Experimental setup and results

In our classification algorithm, the biological motions are bounded by a window of (90×90) pixels. This window allows us to focus on the spatial motion of the human body and limbs in the visual scene. As already mentioned, the motion dynamics of the body and limbs are detected via spatial sampling regions. The radius of the sampling region (mask) is set to 17 pixels with an overlapping of 13 pixels. As a consequence, the number of located masks on the window is $(M_c \times N_c) = (4 \times 4)$ pixels. Since our model for motion estimation drives three speeds and eight directions, the length of the feature vector can be defined by 16 sampling regions $\times$ eight (directions) $\times$ three (speeds)=384 numerical features.

Our classification mechanism uses the cloud of events generated from the bio-inspired model for the motions defined in Section 4.3.2. The feature vectors are extracted within three different temporal windows: 0.10s, 0.15s and 0.20s. Figures 4.4, 4.5 and 4.6 show the confusion matrices of action classification results along 0.10s, 0.15s and 0.20s, respectively. The diagonal boxes of the confusion matrices contain the percentage of correct recognition, while the non-diagonal boxes refer to misclassified results.

The results reveal that the recognition performance is improved when the feature vectors are extracted from a longer temporal window. In other words, the recognition rate based on the feature vector which is extracted along the shortest temporal window (0.10s) leads to more misclassifications than the

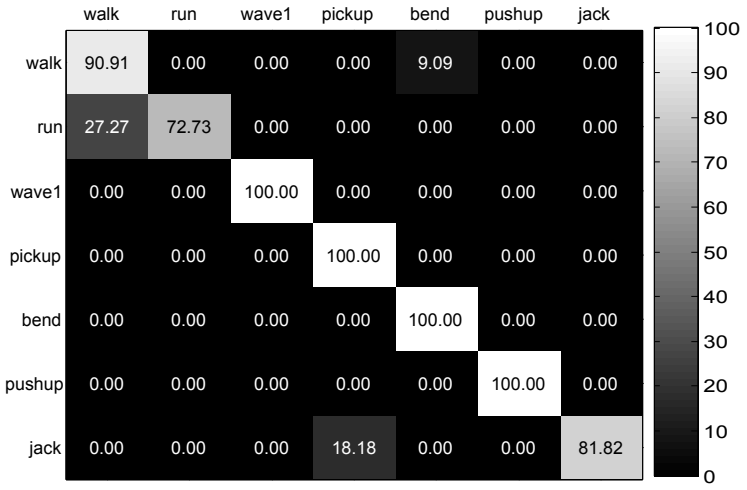


Figure 4.4: Confusion matrix based on feature extraction with temporal integration over 0.10s. The recognition performance is obtained using SVM classifier in which Leave-one-out-procedure is conducted. Diagonal elements of the matrix show the number of correct predictions, while off-diagonal elements show incorrect predictions.

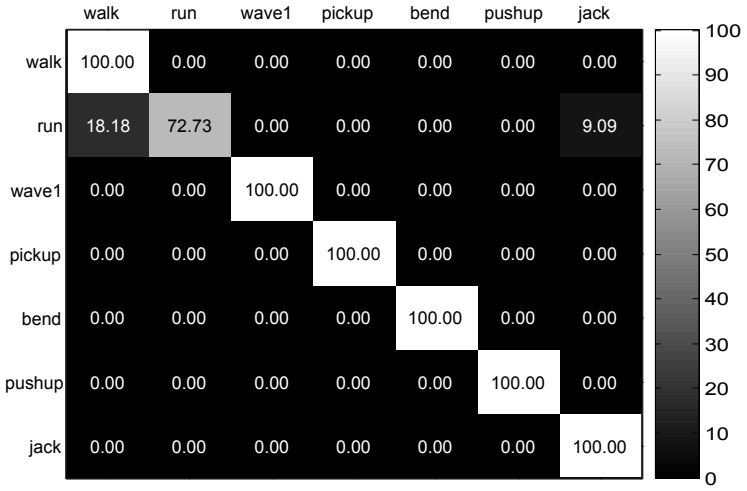


Figure 4.5: Confusion matrix based on feature extraction with temporal integration over 0.15s. The recognition performance is obtained using SVM classifier in which Leave-one-out-procedure is conducted. Diagonal elements of the matrix show the number of correct predictions, while off-diagonal elements show incorrect predictions.



Figure 4.6: Confusion matrix based on feature extraction with temporal integration over 0.20s. The recognition performance is obtained using SVM classifier in which Leave-one-out-procedure is conducted. Diagonal elements of the matrix show the number of correct predictions, while off-diagonal elements show incorrect predictions.

other durations. Figure 4.4 shows one of the walk-class subject classified as a bend-class subject. Moreover, a misclassification occurred between the walk-class and run-class, as well as between the pickup-class and jack-class. This indicates that the feature vector extracted along the 0.10s temporal window is not entirely sufficient to describe the actions. Interestingly, when we used the temporal window of 0.10s, one of the run-class subjects was classified as walk-class. By increasing the time duration of temporal window to 0.15s, the same subject was classified as jack-class (see Figures 4.4 and 4.5). The misclassified results are reduced using the temporal window of 0.20s in which most of the actions are correctly classified except two subjects in the run-class that are classified as walker (see Figure 4.6). This misclassification remains despite of increasing the temporal window ($> 20s$) for feature extraction. This is due to the similarity in the performance of the two actions (walking on the spot and running on the spot) and thus rendering the two subjects perform the running on the spot action close to walking on the spot action.

The results highlight the richness of the extracted features from our model, in which the rates of recognition along the three temporal windows 0.10s, 0.15s and 0.20s are 92.08%, 96.1% and 97.4%, respectively. These results are consistent with the findings of [Schindler et al., 2008] who showed that biological motion can be recognized within a short duration.

In order to show how the test set has an impact on action recognition, we

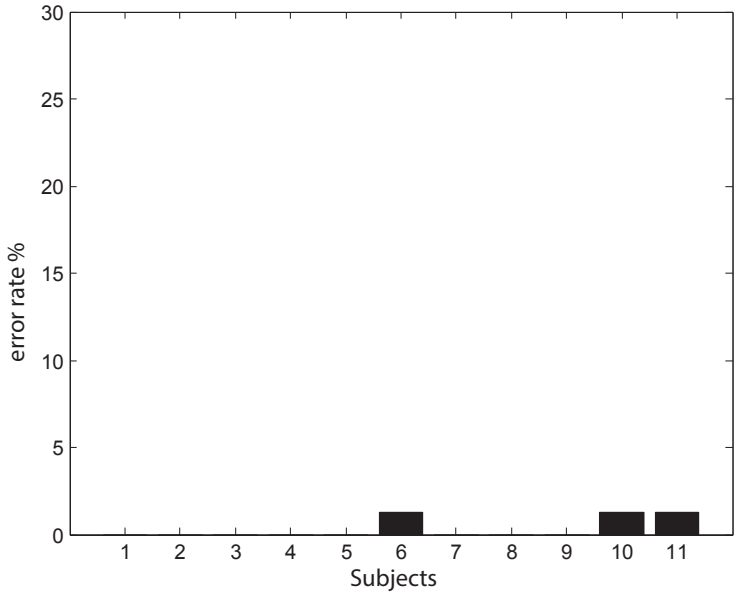


Figure 4.7: The error recognition rate of the test set. The abscissa of the histogram represents the test set in which each subject has an identification number (1 . . . 11). The ordinate of the histogram denotes the error recognition rate of each subject along the whole actions. The test subjects (6, 10, 11) generate an error rate of (2%) which is derived from one misclassification over 49 trails, while 0% error values were generated during the rest of the test set. This indicates that action recognition using a random test subject drives inconclusive decision for recognition performance.

calculated the error rate of recognition results derived from each test subject. Figure 4.7 demonstrates the histogram of the error recognition rate. The abscissa of the histogram represents the test subjects, which are identified by a specific number (1 ... 11). The ordinate of the histogram represents the error rate of the recognition, which is calculated by (error=no. misclassification/no. trails). The result reveals that the test subjects (6, 10, 11) generate an error rate of (2%) which is derived from one misclassification over 49 trails, while 0% error values were generated during the rest of the test set. This highlights that action recognition based on random test subjects is not sufficient to give a thorough conclusion of recognition performance. Our performance of action recognition was thus calculated based on all the possible combinations of test subjects.

4.3.5 Summary and Discussion

In this chapter we have demonstrated the usability of the neural mechanisms defined in Chapters 2 and 3 for classification purposes. The articulated motions are considered for this purpose in which the biological motions are selected as a case study. Action recognition based on a conventional frame-based camera has been taken for granted for many years (e.g., [Bobick and Davis, 2001]; [Efros et al., 2003]; [Escobar and Kornprobst, 2012]) in which an enormous amount of redundant information needs to be processed. Instead, the neuromorphic vision sensor simulates the main processing of the mammalian retina and generates spike-trains of ON/OFF events in response to luminance changes in the visual scene. Our model exploits address-events representation using the DVS128 sensor for data acquisition (see Section 2.5.1). Thus, the spike event input representation introduced therein is generated by recording different human motions using a DVS128 sensor.

The biological motions are characterized as complex motions containing different velocities (directions and speeds) which are generated from the body and limb motions. To extract an essential feature that identifies the human activity, we roughly followed the feature extraction concept suggested by [Escobar and Kornprobst, 2012]. We investigated the richness of the extracted features for action recognition based on our bio-inspired model. The feature vector was derived from the responses of the third stage of our model (area MT) considering direction and speed.

To focus on the human motions in the visual scene, we set a boundary window that surrounds the human movements. As a consequence, only a bounded area of the visual scene is processed. The motion dynamics were detected via spatial sampling regions (4×4) which were placed over the window. We

calculated the average activities that were generated from each mask. These activities were used as a feature vector that describes the corresponding action.

Various aspects have an effect on the performance of the action recognition. The classifier technique has an impact on the recognition rate by attributing the unknown pattern to the right class and minimizing the rate of misclassification. Our algorithm uses nonlinear SVM, which is widely used in the literature and has also shown a reliable classification among different objects (see for e.g., [Jhuang et al., 2007]). On the other hand, the length of the temporal window has an effect on the recognition rate. A very short temporal window (0.10s) leads to misclassifications in the action recognition because this length is not sufficient to capture enough contextual information. On the other hand, by increasing the temporal window to (0.15s and 0.20s), the mechanism can capture more contextual information, and thus the action misclassification rate is decreased.

The training set plays a critical rule in the performance of the classification. Several studies have used a fix training set to evaluate their classification performance (see [Wang et al., 2010, Le et al., 2011]). However we believe that a single training set is not sufficient to give a precise evaluation of the classification algorithm (see [Escobar and Kornprobst, 2012]). This was verified in our algorithm and the histogram in Figure 4.7 shows the dependency of the error distribution on the selected test subject. We thus used all the possible combinations of the test set to evaluate the performance of the classification algorithm where a leave-one-out-procedure is used.

Our approach was verified using different inputs of human motions based on event representation. These motions contain seven actions (walking on the spot, running on the spot, waving one hand, picking up an object from the ground, bending, upper body pushing, and jumping-jack) in which each action was performed by 11 subjects. Our results reveal that the extracted feature derived from the bio-inspired model attained a high level of performance with the average performance of the action recognition tasks being 92.08%, 96.1% and 97.4%.

Our classification mechanism could be extended by extracting the feature vector based on human limbs and ignoring the body motion feature since the human limbs can define an essential feature of an action in the scene. The feature vector could be enriched by adding extra visual properties. These properties could be extracted from higher cortical areas of the motion pathway, such as area MST, or other brain functions of the form pathway. This would enable the model to detect more complicated actions in the visual scene.

Chapter 5

Summary

In this thesis we have proposed a neural model for motion estimation inspired by the visual system of mammals. We have demonstrated how computational mechanisms of visual processing areas can be implemented and how these areas can interact with each other. Since the new technology of event-based sensors simulate the spike-generation functionality of the biological retina, we used a silicon retina (DVS) for visual data acquisition instead of a conventional frame-based sensor. The computation of our model was motivated by neurophysiological and psychophysical studies. These studies have shown that the visual cortex analyzes motion information along a hierarchical architecture composed of several visual areas. These areas interact through bidirectional interconnections in which bottom-up connections pass on the driving visual input from lower areas to higher areas, while top-down connections re-enter higher level representations to the earlier areas via feedback. Motion perception is analyzed along a dorsal pathway starting from the primary visual cortex, area V1. The responses from area V1 are forwarded to subsequent areas such as the middle temporal (MT) and beyond, in which neurons in area V1 are selective for motion direction, while neurons in area MT are selective for motion velocity (speed and direction). Our work thus proposes a bio-inspired visual model in which feedforward and feedback interactions are incorporated. It addresses two main model visual areas with different feature selectivity, V1 and MT. The initial motion is estimated in model area V1 using spatiotemporal filters to locally detect the direction of motion. The responses of area V1 are then weighted and pooled by cells in model area MT which are selective to different velocities, i.e. direction and speed. Such feature selectivity is derived from compositions of activities in the spatio-temporal domain and integrating over larger receptive fields. In order to account for the bidirectional coupling of cortical areas, we mapped properties of the feature selectivity in both areas for feedback processing. We integrated the responses over different speeds along

a particular preferred direction. In order to regulate the overall activation of single neurons against a pool of neighboring cells, we incorporated a competitive stage that operates upon the spatial as well as the feature domain. The results of the optical flow estimation were analyzed using synthetic ground truth data. We tested our model using different kinds of stimuli moving in different directions. The results showed an improvement in the flow estimation in area MT and modulated V1, compared with the initial motion estimation in area V1. In order to examine the performance of our model for classification tasks, we selected action recognition as a case study using different biological movements with satisfactory results. The following section introduces the main contributions of our work.

5.1 Main contributions

The model is mainly about motion detection, integration and the signaling along bottom-up as well as top-down streams. The input is derived from event-streams generated by a new camera technology. For the motion detection and integration, we modeled two areas of the visual cortex, V1 and MT. The bottom-up signals (V1$\rightarrow$ MT) are predominantly driven by spatio-temporal features (which do not distinguish between different speeds). The top-down signals (MT $\rightarrow$ V1), on the other hand, are driven by signals derived from representations that are sensitive to direction as well as speed. The questions here are how is it possible to build a model that is consistent with the spike generation of the silicon retinas and how can the higher area signals interact with the signals of the lower area where both are fed by different feature domains? Our model proposes possible solutions to these questions, which can be divided into four main contributions:

- We adapted the spatio-temporal filters of the type suggested by findings of [Adelson and Bergen, 1985] to be consistent with the address-events representation. In the model of Adelson-Bergen, the authors suggested to utilize temporal gamma functions of different duration in order to accomplish temporal smoothing and differentiation, leading to a temporally biphasic response shape. In order to transcribe this functionality to the spike-trains output of the DVS sensor, we make use of the following approximation: The biphasic Adelson-Bergen temporal filters can be decomposed into a convolution of numerical difference kernel (to approximate a first-order derivative operation) with a temporal smoothing filter. The event-based sensor already operates by generating discrete events based on changes, i.e. temporal derivatives, in the input signal. For that

reason, we employ temporal smoothing filters which are calculated from the integral of Adelson-Bergen temporal filters (see Equation (2.10)) and convolve them with the input stream of events to obtain scaled versions of temporally smoothed derivatives of the input luminance function. To simplify the integral operation, we suggest to reconstruct Adelson-Bergen gamma functions by combining two temporally offset Gaussian functions. Here, the integration results of the Gaussian combinations generate new temporal filters that consistent with the spike-generation of the DVS sensor.

- We proposed a new mechanism to simulate the functionality of the cortical area MT. The suggested neural mechanisms of area MT is motivated by principle findings of neuroscientists studies. The localized measures from area V1 are integrated in area MT [Born and Bradley, 2005] through larger RFs which are roughly ten times the RFs of V1 cells [Maunsell and van Essen, 1987, Albright and Desimone, 1987]. In our model, we integrated the responses of the spatio-temporal filters through RFs, or filters, which are larger in their size by up to an order of magnitude. However, the limited spatial resolution of the DVS sensor (128×128) [Lichtsteiner et al., 2008] handicaps us from increasing the size of the MT filters to ten time the size of V1 filters. For that reason, we set the filters sizes with range of (V1:MT 1:$\geq$3). This problem is however not a conceptual one of our approach and will most likely be solved with future versions of the DVS sensor. The geometry of the RFs in area MT has been investigated by [Born, 2000] in which two types of RFs profiles are defined. The first type is classical receptive field (CRF) which responds best to wide field motion. While the second type is center-surround cell which is sensitive to motion contrast. In our modeling, we focused on CRF cells in which Gaussian weighting function has been used to describe the RFs profiles. This function has been used in many studies to describe the neural RF model (see e.g., [Escobar and Kornprobst, 2012, Tschechne et al., 2014a, Bayerl and Neumann, 2004]). The neural functionality of MT cells is modeled to be selective to different directions and speeds which is consistent with the physiological findings [Albright, 1984, Felleman, 1984, Lagae et al., 1993]. Here, we integrate the early motion responses of area V1 by utilizing circular RFs weighting functions with Gaussian profile. Such integration operates at a much larger spatial context to properly integrate localized responses of area V1. As a consequence, the uncertain flow estimation will be enhanced in area MT. In order to equip MT cells with different speed selectivity, we suggest to incorporate elongated RFs of Gaussian weighting function in our model.

These RFs are oriented in the spatial-temporal domain which, in turn, enables the integration strategy to increase the speed selectivity of MT cells to different speeds (slow, mid and fast motions). The orientation in space-time domain with different angles encodes the speed of the input stimulation in which high oriented angles represent high-speed detectors, while small oriented angles represent slow-speed detectors.

- To enhance the initial motion estimation in area V1 and reduce ambiguity in the estimated motion, top-down feedback signals from area MT are fed back and re-entered in area V1. Along the feedforward path different feature representations of the spatio-temporal input are generated through hierarchically organized filter processes. The specific features extracted are spatio-temporal frequency (in V1) and velocity (direction, speed) at spatial locations (in MT), respectively. In order to generate a proper feedback signals from model area MT that can combine with the spatio-temporal responses in V1, we integrated the responses over different speeds along a particular preferred direction. Such matching allows top-down feedback signals from area MT to to be compatible with the feature property of area V1.

- We incorporated the response normalization in our model to achieve balance activities of individual neurons in areas V1 and MT. The interaction between the normalization of responses and the enhancement activities via feedback projection establishes the dynamics of visual cortical processing. Following the suggestion of [Bouecke et al., 2011] and theoretical studies (e.g., [Brosch and Neumann, 2014]) we carried out the normalization process using model neuronal activation which is described by [Grossberg, 1988] as gradual changes of the membrane potential. Activity normalization of model area V1 and MT is computed by realizing a slightly simplified version of the scheme described in [Brosch and Neumann, 2014] and solve the normalization interaction at equilibrium, namely evaluating the state response for $\frac{dv(t)}{dt} = 0$. Here, the normalization process is carried out in spatial domain by calculating the pool of activities of individual neuron over a circular spatial neighborhood. In addition, we normalize neuron activities in the feature domain. Since different properties are derived from both visual areas, V1 (directions and spatio-temporal selectivity) and MT (speed and directions), we normalized responses of area V1 by averaging activities of each cell over all directions. In area MT, on the other hand, we normalized the responses by averaging activities for individual directions by integrating the activities of cells over different speeds.

5.2 Relevant publications

Abdul-Kreem, L. I. and Neumann, H. (2016). Estimating visual motion using an event-based artificial retina. Series of CCIS Communications in Computer and Information Science. Computer Vision, Imaging and Computer Graphics Theory and Applications. Springer International Publishing Switzerland 2016

Abdul-Kreem, L. I. and Neumann, H. (2015.). Neural mechanisms of cortical motion computation based on a neuromorphic sensory system. PLOS ONE, DOI:10.1371/journal.pone.0142488. November 10.

Abdul-kreem, L. I. and Neumann, H. (2015). Bioinspired model for motion estimation using address event representation. 10th International Conference on computer vision theory and application,VISIGRAPP, (Proc. 10th Int'l. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2015; March 11-14, Berlin, Germany, 2015).

Tschechne, S., Brosch, T., Sailer, R., Egloffstein, N., **Abdul-kreem, L. I.**, and Neumann, H. (2014). On event-based motion detection and integration. Proc. 8th Int'l Conf. on Bio-inspired Information and Communication Technologies, BICT 2014, Dec.1-3, Boston, MA, USA, ACM digital library.

Bibliography

[Adelson and Bergen, 1985] Adelson, E. and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):90–105.

[Adelson and Movshon, 1982] Adelson, E. and Movshon, J. (1982). Phenomenal coherence of moving visual pattern. *Nature*, 300(5892):523–525.

[Albright, 1984] Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque. *Journal of Neurophysiology*, 52(6):1106–1129.

[Albright and Desimone, 1987] Albright, T. D. and Desimone, R. (1987). Local precision of visutopic organization in the middle temporal area (mt) of the macaque. *Experimental Brain Research*, 65(3):582–592.

[Azimuth, 2016] Azimuth (22.3.2016). *https://johncarlosbaez.wordpress.com/2012/05/21/symmetry-and-the-fourth-dimension-part-1/(blog).*

[Barron et al., 1994] Barron, J., Fleet, D., and S., B. (1994). Performance of optical flow techniques. *The International Journal of Computer Vision*, 12(1):43–77.

[Bastos et al., 2012] Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., Friston, K. J. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

[Bastos et al., 2014] Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J., Oostenveld, R., Dowdall, J. R., Weerd, P. D., Kennedy, H., and Fries, p. (2014). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Available at http://biorxiv.org/content/early/2014/05/06/004804*.

[Bauckhage et al., 2009] Bauckhage, C., Tsotsos, J. K., and Bunn, F. E. (2009). Automatic detection of abnormal gait. *Image and vision computing*, 27(1-2):108–115.

[Bayerl and Neumann, 2004] Bayerl, P. and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16(10):2041–2066.

[Bear et al., 2007] Bear, M. F., Connors, B. W., and Paradiso, M. A. (2007). *Neuroscience: Exploring the Brain, Third Edition*. Lippincott Williams & Wilkins.

[Beauchemin and Barron, 1995] Beauchemin, S. S. and Barron, J. L. (1995). The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–466.

[Beintema and Lappe, 2002] Beintema, J. and Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the USA*, 99(8):5661–5663.

[Benosman et al., 2011] Benosman, R., Ien, S., Rogister, P., and Posch, C. (2011). Asynchronous event-based hebbian epipolar geometry. *IEEE Trans. Neural Networks and Learning Systems*, 22(11):1723–1734.

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

[Blake and Shiffrar, 2007] Blake, R. and Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, 58:47–73.

[Blomfield, 1974] Blomfield, S. (1974). Arithmetical operations performed by nerve cells. *Brain Research*, 69(1):115–124.

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. On pattern Analysis and Machine Intelligence*, 23(3):257–267.

[Born, 2000] Born, R. T. (2000). Centersurround interactions in the middle temporal visual area of the owl monkey. *Neurophysiology*, 84(5):2658–2669.

[Born and Bradley, 2005] Born, R. T. and Bradley, D. C. (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, Article ID 781561, 22 pages.*

[Bower and Beeman, 1998] Bower, J. M. and Beeman, D. (1998). *Compartmental modeling. In The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: Springer-Verlag.

[Brandli et al., 2012] Brandli, C., Berner, R., Yang, M., Liu, S. C., and Delbruck, T. (2012). A 240 $\times$ 180 130db 3μ latency global shutter spatiotemporal vision sensor. *IEEE Journal of solid-state circuits*, 49(10):2333–2341.

[Brosch and Neumann, 2014] Brosch, T. and Neumann, H. (2014). Computing with a canonical neural circuits model with pool normalization and modulating feedback. *Neural Computation*, 26(12):2735–2789.

[Brosch et al., 2015] Brosch, T., Tschechne, S., and Neumann, H. (2015). On event-based optical flow detection. *Frontiers in Neuroscience*, 137(9).

[Brox et al., 2004] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. *In Proc.8th European Conference on Computer Vision, Springer LNCS 3024, T. Pajdle and J. Matas(Eds), (Prague, Czech Republic)*, pages 25–36.

[Bush and Sejnowski, 1993] Bush, P. C. and Sejnowski, T. J. (1993). Reduced compartmental models of neocortical pyramidal cells. *Journal of Neuroscience Methods*, 46(2):159–166.

[Camunas-Mesa et al., 2014] Camunas-Mesa, L. A., T., S.-G., Leng, 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).

[Caplovitz et al., 2006] Caplovitz, G., Hsieh, P., and Tse, P. (2006). Mechanisms underlying the perceived angular velocity of a rigidly rotating object. *Vision Reseach.*, 46(18):2877–2893.

[Carandini and Heeger, 1994] Carandini, M. and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264(5163):1333–1336.

[Carandini and Heeger, 2012] Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13:51–62.

[Carandini et al., 1999] Carandini, M., Heeger, D. J., and Movshon, J. A. (1999). Linearity and gain control in v1 simple cells. *in Model sof Cortical Circuits , Volume 13 of Cerebral Cortex, eds P. S. Ulinski, E. G.Jones, and A. Peters (New York, NY: Springer US, Kluwer Academic/Plenum Publishers)*, 264:401–443.

[Casile and Giese, 2003] Casile, A. and Giese, M. (2003). Role of motion and form in biological motion recognition. *ICANN/ICONIp, LNCS 2714*, pages 854–862.

[Casile and Giese, 2005] Casile, A. and Giese, M. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, 5(4):348–360.

[Challinor and Mather,] Challinor, K. L. and Mather, G. A motion-energy model predicts the direction discrimination and mae duration of two-stroke apparent motion at high and low retinal illuminance. *Vision Research*, 50:1109–1116.

[Chalupa and Werner, 2004] Chalupa, L. M. and Werner, J. S. (2004). *The visual neurosciences*. The MIT press, Cambridge, Massachusetts USA.

[Chey et al., 1997] Chey, J., Grossberg, S., and Mingolla, E. (1997). Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction. *Journal of the Optical Society of America A*, 14(10):2570–2594.

[Chey et al., 1998] Chey, J., Grossberg, S., and Mingolla, E. (1998). Neural dynamics of motion processing and speed discrimination. *Vision Research*, 38(18):2769–2786.

[Churchland et al., 2005] Churchland, M. M., Priebe, N. J., and Lisberger, S. G. (2005). Comparison of the spatial limits on direction selectivity in visual area mt and v1. *Journal of Neurophysiology*, 93(3):1235–1245.

[Conradt et al., 2009] Conradt, J., Berner, R., Cook, M., and Delbruck, T. (2009). An embedded aer dynamic vision sensor for low-latency pole balancing. *5th IEEE Workshop on Embedded Computer Vision (in conjunction with ICCV)*, pages 1–6.

[Conway and Livingstone, 2003] Conway, B. and Livingstone, M. (2003). Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque v1. *Journal of Neurophysiology*, 89(5):2726–2742.

[Crick and Koch, 1998] Crick, F. and Koch, C. (1998). Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature*, 391(6664):245–250.

[Crouzet and Cauchoix, 2011] Crouzet, S. M. and Cauchoix, M. (2011). When does the visual system need to look back? *Neuroscience*, 31(24):8706–8707.

[Cutting and kozolowski, 1977] Cutting, J. E. and kozolowski, L. T. (1977). Recognizing friends by their walk: gait perception without fariamility cues. *Bulletin of the psychonomic Society*, 9(5):353–356.

[Datta et al., 2011] Datta, A., Sheikh, Y., and Kanade, T. (2011). Linearized motion estimation for articulated plane. *IEEE Transactions on pattern analysis and machine intelligence*, 33(4):780–793.

[Daugman, 1985] Daugman, J. (1985). Uncertainty relation for resolution , spatial frequency, and orientation optimized by 2d visual cortical filters. *Journal of the Optical Society of America A 2*, 2(7):1160–1169.

[Daugman, 1988] Daugman, J. G. (1988). Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on acoustics, speech, and signal processing*, 36(7):1169–1179.

[Dayan and Abbot, 2001] Dayan, P. and Abbot, L. F. (2001). Theoretical neuroscience. *MIT Press,Cambridge, Mass, USA,*.

[De Valois et al., 2000] De Valois, R. L., Cottaris, N. P., Mahon, L. E., Elfar, S. D., and Wilson, J. A. (2000). Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vision Research*, 40(27):3685–3702.

[De Valois et al., 1982] De Valois, R. L., Yund, E. W., and Kepler, N. K. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Res*, 22(5):531–544.

[De Valoisa et al., 2000] De Valoisa, R., Cottarisb, N. P., Mahonb, L. E., Elfara, S. D., and Wilsona, J. A. (2000). Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vision Research*, 40(27):3685–3702.

[DeAngelis et al., 1995] DeAngelis, G., Ohzawa, I., and Freeman, R. (1995). Receptive-field dynamics in the central visual pathways. *Trends Neurosci*, 18(10):451–458.

[Delbruck, 2008] Delbruck, T. (2008). Frame-free dynamic digital vision. *Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society, Univ. of Tokyo, Mar. 6-7*, pages 21–26.

[Delbruck, 2012] Delbruck, T. (2012). Fun with asynchronous vision sensors and processing. *A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, © Springer-Verlag Berlin Heidelberg*, pages 506–515.

[Delbrück and Lang, 2013] Delbrück, T. and Lang, M. (2013). Robotic goalie with 3 ms reaction time at 4 % cpu load using event-based dynamic vision sensor. *Frontiers in Neuroscience*, 7(223):1–7.

[Delbruck and Lichtsteiner, 2008] Delbruck, T. and Lichtsteiner, P. (2008). Fast sensory motor control based on event-based hybrid neuromorphic-procedural system. *IEEE International Symposiom on circuit and system*, pages 845 – 848.

[Delbruck et al., 2010] Delbruck, T., Linares-Barranco, B., Culurciello, E., and Posch, C. (2010). Activity-driven, event-based vision sensors. *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 2426–2429.

[Derrington and Webb, 2004] Derrington, A. M. and Webb, B. S. (2004). Visual system: How is the retina wired up to the cortex? *Current Biology*, 14(1):R14–R15.

[Desimone, 1998] Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, 353(1373):1245–1255.

[Desimone and Duncan, 1995] Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222.

[DeValois and Cottaris, 1998] DeValois, R. L. and Cottaris, N. P. (1998). Inputs to directionally selective simple cells in macaque striate cortex. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14488–14493.

[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(5):1465–1469.

[Drulea and Nedevschi, 2013] Drulea, M. and Nedevschi, S. (2013). Motion estimation using the correlation transform. *IEEE Transaction on Image Processing*, 22(8):1057–7149.

[Edelman, 1993] Edelman, G. M. (1993). Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115125.

[Efros et al., 2003] Efros, A., Berg, A., Mori, G., and Malik, J. (2003). Recognizing action at a distance. *Proceedings of the 9th International Conference on Computer Vision, Nice, France, Oct 1316*, pages 726–734.

[Emerson et al., 1987] Emerson, R., Citron, M., Vaughn, W., and Klein, S. (1987). Nonlinear directionally selective subunits in complex cells of cat striate cortex. *Journal of Neurophysiology*, 58(1):33–65.

[Emerson et al., 1992] Emerson, R. C., Bergen, J. R., and Adelson, E. H. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, 32(2):203–218.

[Escobar, 2009] Escobar (2009). Bio-inspired models for motion estimation and analysis: Human action recognition and motion integration. *Ph.D. thesis, University of Nice-Sophia Antipolis, Specialized in Control, Signal and Image Processing, Nov. 27.*

[Escobar and Kornprobst, 2012] Escobar, M. J. and Kornprobst, P. (2012). Action recognition via bio-inspired feature: the richness of center-surround interaction. *Computer Vision and Image Understanding*, 116(5):593–605.

[Fahle and Poggio, 1981] Fahle, M. and Poggio, T. (1981). Visual hyperacuity: spatiotemporal interpolation in human vision. *Proceedings of the Royal Society of London*, 213(1193):451–477.

[Felleman and Van Essen, 1991] Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

[Felleman, 1984] Felleman, D. J., K. J. H. (1984). Receptive-field properties of neurons in middle temporal visual area (mt) of owl monkeys. *J. Neurophysiol*, 52(3):488–513.

[Fleet and Jepson, 1990] Fleet, D. and Jepson, A. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104.

[Fleet and Jepson, 1984] Fleet, D. J. and Jepson, A. D. (1984). A cascaded filter approach to the construction of velocity selective mechanisms. *Technical Report RBCV-TR-84-6, Department of Computer Science, University of Toronto.*

[Fleet and Weiss, 2006] Fleet, D. J. and Weiss, Y. (2006). *Handbook of Mathematical Models in Computer Vision: Chapter 15, Optical flow estimation.* Springer Science and Business media, Inc.

[Friston and Büchel, 2000] Friston, K. J. and Büchel, C. (2000). Attentional modulation of effective connectivity from v2 to v5/mt in humans. *Proc. Natl. Acad. Sci. USA 97*, pages 7591–7596.

[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.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers*, 93(26):429–460.

[Gibson et al., 2014] Gibson, T. A., Heath, S., Quinn, R. P., Lee, A. H., Arnold, J., Sonti, T. S., Whalley, A., Shannon, G. P., Song, B. T., Henderson, J. A., and Wiles, J. (2014). Event-based visual data sets for prediction tasks in spiking neural networks. *In ICANN 2014, LNCS 8681. Springer Int. Pub. Switzerland*, pages 635–642.

[Giese and Poggio, 2003] Giese, M. A. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movement. *Nature Reviews Neuroscience*, 4:179–192.

[Girard and Bullier, 1989] Girard, P. and Bullier, J. (1989). Visual activity in area v2 during reversible inactivation of area 17 in the macaque monkey. *J. Neurophysiol*, 62(6):1287–1302.

[Girosi et al., 1989] Girosi, F., Verri, A., and Torre, V. (1989). Constraints for the computation of optical flow. *Proc. IEEE Workshop on Visual Motion. Irvine, CA, 20-22 Mar*, pages 116–124.

[Gorelick and Blank, 2007] Gorelick, L. and Blank, M. (2007). Action as space-time shapes. *IEEE Transactions on pattern Analysis and machine intelligence*, 29(12):I:635–I640.

[Green et al., 2004] Green, W. E., Oh, P. Y., and Barrows, G. (2004). Flying insect inspired vision for autonomous aerial robot maneuvers in near-earth environments. *Proceedings of the 2004 IEEE International Conference on robotics and Automation*, 3:2347–2352.

[Grossberg, 1980] Grossberg, S. (1980). How does the brain build a cognitive code? *Psychol. Rev.*, 87(1):1–51.

[Grossberg, 1988] Grossberg, S. (1988). Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1(1):17–61.

[Grossberg, 1994] Grossberg, S. (1994). 3d vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, 55(1):48–121.

[Grossberg et al., 2001] Grossberg, S., Mingolla, E., and Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research*, 41(19):2521–2553.

[Grzywacz and Yuille, 1990] Grzywacz, N. and Yuille, A. (1990). A model for the estimate of local image velocity by cells on the visual cortex. *Proc R Soc Lond B Biol Sci.*, 239(1295):129–161.

[Heeger, 1988] Heeger, D. J. (1988). Optical flow using spatiotemporal filters. *Intl. J. Comp. Vis.*, 1(4):279–302.

[Heeger, 1992] Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197.

[Herz et al., 2006] Herz, A., Gollisch, T., Machens, C., and Jaeger, D. (2006). Modeling single-neuron dynamics and computations: a balance of detail and abstraction. *Science*, 314(5796):80–85.

[Hildreth and Koch, 1987] Hildreth, E. C. and Koch, C. (1987). The analysis of visual motion: From computational theory to neuronal mechanisms. *Annual Review of Neuroscience*, 10:477–533.

[Hirai and Hiraki, 2006] Hirai, M. and Hiraki, K. (2006). Neural dynamics for biological motion perception. *In: F. J. Chen (ed.): Trends in Brain Mapping Research. Nova Sciences Publishers*, pages 85–116.

[Hodgkin, 1937a] Hodgkin, A. L. (1937a). Evidence for electrical transmission in nerve, part i. *Journal of Physiology*, 90(2):183–210.

[Hodgkin, 1937b] Hodgkin, A. L. (1937b). Evidence for electrical transmission in nerve, part ii. *Journal of Physiology*, 90(2):211–32.

[Hodgkin and Huxley, 1952a] Hodgkin, A. L. and Huxley, A. F. (1952a). The components of membrane conductance in the giant axon of loligo. *J. Physiol*, 116(4):473–496.

[Hodgkin and Huxley, 1952b] Hodgkin, A. L. and Huxley, A. F. (1952b). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol*, 117(4):500–544.

[Horn and Schunck, 1981] Horn, B. and Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

[Hubel, 1995] Hubel, D. H. (1995). *Eye, Brain, and Vision*. Henry Holt and Company, 2nd edition.

[Hubel and Wiesel, 1959] Hubel, D. H. and Wiesel, T. N. (1959). Receptive field of single neurons in the cat's striate cortex. *J. Physiol*, 148(3):574–591.

[Hubel and Wiesel, 1960] Hubel, D. H. and Wiesel, T. N. (1960). Receptive fields of optic nerve fibres in the spider monkey'. *J. Physiol*, 154(3):572–80.

[Hubel and Wiesel, 1961] Hubel, D. H. and Wiesel, T. N. (1961). Integrative action in the cat's lateral geniculate body. *J. Physiol*, 155(2):385–398.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol*, 160(1):106–154.

[Hubel and Wiesel, 1968] Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol*, 195(1):215–43.

[Jhuang et al., 2007] Jhuang, H., Serre, T., and Wolf, L. Poggio, T. (2007). A biologically inspired system for action recognition. *Processing of the 11th International Conference on Computer Vision*, pages 1–8.

[Johansson, 1973] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211.

[Johnson and Burkhalter, 1997] Johnson, R. and Burkhalter, A. (1997). A polysynaptic feedback circuit in rat visual cortex. *J. Neurosci.*, 17(18):7129–7140.

[Juhansson, 1976] Juhansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research*, 38(4):379–393.

[Kandel et al., 2000] Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (2000). *Principles of Neural Science*. McGraw-Hill, 4th Edition.

[Kara and Reid, 2003] Kara, P. and Reid, R. C. (2003). Efficacy of retinal spikes in driving cortical responses'. *The Journal of Neuroscience*, 23(24):8547–8557.

[Koch, 1998] Koch, C. (1998). *Biophysics of computation: information processing in single neurons*. Oxford Univ. Press, New York.

[Koch and Davis, 1994] Koch, C. and Davis, J. L. (1994). *Large-scale neuronal theories of the brain.* Massachusetts Institute of Technology, A Bradford Book, The MIT press, Cambridge, Massachusetts, London, England.

[Koivisto et al., 2011] Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., and Salminen-Vaparanta, N. (2011). Recurrent processing in v1/v2 contributes to categorization of natural scenes. *J. Neurosci.*, 31(7):2488–2492.

[König and Luksch, 1989] König, P. and Luksch, H. (1989). Rockland ks, virga a (1989) terminal arbors of individual feedback axons projecting from area v2 to v1 in the macaque monkey: A study using immunohistochemistry of anterogradely transported phaseolus vulgaris- leucoagglutinin. *J. Comp. Neurol.*, 285(1):54–72.

[König and Luksch, 1998] König, P. and Luksch, H. (1998). Active sensing: Closing multiple loops. *Z Naturforsch C.*, 53(7-8):542–549.

[Kouh and Poggio, 2008] Kouh, M. and Poggio, T. (2008). A canonical neural circuit for cortical non-linear operations. *Neural Comput*, 20(6):1427–1451.

[Kuffler, 1953] Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68.

[Kuznetsov et al., 2006] Kuznetsov, A. S., Kopell, N. J., and Wilson, C. J. (2006). Transient high-frequency firing in a coupled-oscillator model of the mesencephalic dopaminergic neuron. *Journal of Neurophysiology*, 95(2):932–947.

[Lagae et al., 1993] Lagae, L., Raiguel, S., and Orban, G. A. (1993). Speed and direction selectivity of macaque middle temporal neurons. *J Neurophysiol*, 69(1):19–39.

[Lamme et al., 1998] Lamme, V. A., Supér, H., and Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Curr. Opin. Neurobiol*, 8(4):529–535.

[Laptev, 2005] Laptev, I. (2005). On apace-time interest point. *International Journal of Computer vision*, 64(2/3):107–123.

[Laptev et al., 2007] Laptev, I., Caputo, B., Schuldt, C., and Lindeberg, T. (2007). Local velocity adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3):207–229.

[Larkum, 2013] Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *TrendsNeurosci*, 36(3):141–151.

[Layher et al., 2014a] Layher, G., Giese, M., and Neumann, H. (2014a). Learning representations of animated motion sequences-a neural model. *Proc. 35th Annual Conf. of the Cognitive Science Society CogSci*, pages 170–182.

[Layher et al., 2014b] Layher, G., Schrodt, F., Butz, M. V., and Neumann, H. (2014b). Adaptive learning in a compartmental model of visual cortex-how feedback enables stable category learning and refinement. *Frontiers in Psychology*, 5.

[Le et al., 2011] Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3361–3368.

[Lee and Delbruck, 2014] Lee, J. H. and Delbruck, T. (2014). Real-time gesture interface based on event-driven processing from stereo silicon retinas. *IEEE transaction on neural networks and learning systems*, 25(12):2250–2263.

[Lee et al., 2012] Lee, J. H., Delbruck, T., Park, P., Pfeiffer, M., Shin, C., Ryu, H., and Kang, B. (2012). Gesture based remote control using stereo pair of dynamic vision sensors. *In IEEE Intl. Symp. on Circuits and systems(ISCAS)*, pages 741–745.

[Lichtsteiner et al., 2008] Lichtsteiner, P., Posch, C., and Delbruck, T. (2008). A 128 $\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576.

[Linde and Lindeberg, 2004] Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histogram of higher dimensionality. *Proc Int. Conf. on Pattern Recognition, Cambridge, UK*, pages II:1–II:6.

[Litzenberger et al., 2006a] Litzenberger, M., Belbachir, A. N., Donath, N., Gritsch, G., Garn, H., Kohn, B., Posch, C., and Schraml, S. (2006a). Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor. *Proceeding 6th IEEE Intelligent Transportation Systems Conference Toronto, Canada*, pages 17–20.

[Litzenberger et al., 2006b] Litzenberger, M., Posch, C., Bauer, D., Belbachir, A. N., Schon, P., Kohn, B., and Garn, H. (2006b). Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. *IEEE DSPW, 12th - Signal Processing Education Workshop*, pages 173–178.

[Liu and Delbruck, 2010] Liu, S. and Delbruck, T. (2010). Neuromorphic sensory systems. *Neurobiology*, 20(3):288–295.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with and application to stereo vision. *In Proceedings of Imaging Understanding Workshop*, pages 121–130.

[Lyu and Simoncelli, 2009] Lyu, S. and Simoncelli, E. (2009). Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, 21(6):1485–1519.

[Mahowald, 1992] Mahowald, M. A. (1992). Vlsi analogs of neuronal visual processing: a synthesis of form and function. *Ph.D. dissertation, California Institute of Technology, 12 May.*

[Mahowald, 1994] Mahowald, M. A. (1994). An analog vlsi system for stereoscopic vision. *Springer International Series in Engineering and Computer Science, Kluwer, Boston*, 265.

[Mahowald and Mead, 1991] Mahowald, M. A. and Mead, C. (1991). The silicon retina. *Sci. Amer.*, 264:76–82.

[Markov et al., 2014] Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., and Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.*, 522(1):225–259.

[Maunsell and van Essen, 1987] Maunsell, J. H. R. and van Essen, D. C. (1987). Topographic organization of the middle temporal visual area in the macaque monkey: representational biases and the relationship to callosal connections and myeloarchitectonic boundaries. *J. Comp. Neurol.*, 266(4):535–555.

[Michels et al., 2005] Michels, L., Lappe, M., and Vaina, L. (2005). Visual areas involved in the perception of human movement from dynamic analysis. *Brain Imaging*, 16(10):1037–1041.

[Mikami et al., 1986] Mikami, A., Newsome, W., and Wurtz, R. (1986). Motion selectivity in macaque visual cortex.ii. spatiotemporal range of directional interactions in mt and v1. *Journal of Neurophysiology*, 55(6):1328–1339.

[Milgram et al., 2006] Milgram, J., Cheriet, M., and Sabourin, R. (2006). "one against all": which one is better for handwriting recognition with svms? *in 10th International Workshop on Frontiers in Handwriting Recognition, La Baule, France*, pages 492–497.

[Mishkin et al., 1983] Mishkin, M., Ungerleider, L., and Macko, K. (1983). Object vision and spatial vision: two central pathways. *Trend in Neuroscince*, 6:414–417.

[Moeslund and Granum, 2001] Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268.

[Movshon et al., 1985] Movshon, J. A., Adelson, E. H., Gizzi, M. S., and Newsome, W. T. (1985). *The analysis of moving visual patterns. In: Pattern recognition mechanisms.* (Chagas C, Gattass R, Gross C, eds), Rome: Vatican Press. Pont Acad Sci Scr Varia 54: 117-151.

[Movshon et al., 1978a] Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978a). Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, 283:79–99.

[Movshon et al., 1978b] Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology*, 283:53–77.

[Mumford, 1991] Mumford, D. (1991). On the computational architecture of the neocortex:i. the role of the thalamo-cortical loop. *Biol Cybern*, 65:135–145.

[Murray and Basu, 1994] Murray, D. and Basu, A. (1994). Motion tracking with an active camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):449–459.

[Nagel, 1987] Nagel, H. H. (1987). On the estimation of optical flow: Relation between different approaches and some new results. *Artificial Intelligence*, 33(3):299–324.

[Neumann and Sepp, 1999] Neumann, H. and Sepp, W. (1999). Recurrent v1-v2 interaction in early visual boundary processing. *Biol.Cybern.*, 81:425–444.

[Ni et al., 2011] Ni, Z., Pacoret, C., Benosman, R., Ieng, S., and Regnier, S. (2011). Asynchronous event-based high speed vision for microparticle tracking. *Journal of Microscopy*, 43(2):1365–2818.

[Niebles and Fei-Fei, 2007] Niebles, J. and Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. *Proceeding 7th IEEE Computer Vision and Pattern Recognition, 17-22 June*, pages 207–229.

[Nilsback and Caputo, 2004] Nilsback, M. E. and Caputo, B. (2004). Cue integration through discriminative accumulation. *Proceeding IEEE computer vision and pattern recognition, 27 June-2 July*, pages II.578–II:585.

[Orban de Xivry et al., 2010] Orban de Xivry, J.-J., Coppe, S., Lefvre, P., and Missal, M. (2010). Biological motion drives perception and action. *Journal of Vision*, 10(2):1–11.

[Pack et al., 2006] Pack, C., Conway, B., Born, R., and Livingstone, M. (2006). Spatiotemporal structure of nonlinear subunits in macaque visual cortex. *Journal of Neuroscience*, 26(3):893–907.

[Perrone and Krauzlis, 2002] Perrone, J. and Krauzlis, R. (2002). Simulating the time course of mt neuron responses with a model based on v1 neuron properties, [abstract]. *Journal of Vision*, 2(7):38.

[Perrone, 2004] Perrone, J. A. (2004). A visual motion sensor based on the properties of v1 and mt neurons. *Vision Research*, 44(15):1733–1755.

[Perrone and Thiele, 2001] Perrone, J. A. and Thiele, A. (2001). Speed skills: measuring the visual speed analyzing properties of primate mt neurons. *nature neuroscience*, 4(5):526–532.

[Perrone and Thiele, 2002] Perrone, J. A. and Thiele, A. (2002). A model of speed tuning in mt neurons. *Vision Research*, 42(8):1035–1051.

[Pinto and Shiffar, 1999] Pinto, J. and Shiffar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica*, 102(2-3):293–318.

[Poggio and Reichardt, 1973] Poggio, T. and Reichardt, W. (1973). Considerations on models of movement detection. *Kybernetik*, 13(4):223–227.

[Poort et al., 2012] Poort, J., Raudies, F., Wannig, A., Lamme, V. A., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas v1 and v4 of the visual cortex. *Neuron*, 75(1):143–156.

[Priebe et al., 2006] Priebe, N. J., Lisberger, S. G., and Moshon, A. J. (2006). Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *Journal of Neuroscience*, 26(11):2941–2950.

[Rao and Ballard, 1999] Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex:a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2:79–87.

[Reichardt, 1961] Reichardt, W. (1961). *Autocorrelation, a principle for the evaluation of sensory information by the central nervous system.* in Sensory Communication, W. A. Rosenblith, ed. (Wiley, New York).

[Reynolds and Heeger, 2009] Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61(2):168–185.

[Ringach, 2002] Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Neurophysiol*, 88(1):455–463.

[Rodman and Albright, 1989] Rodman, H. R. and Albright, T. D. (1989). Single-unit analysis of patten-motion selective properties in the middle temporal visual area (mt). *Exp Brain Res*, 75(1):53–64.

[Roelfsema et al., 2002] Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., and Bosch, H. (2002). Figure-ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.*, 14(4):525–537.

[Rolls and Stringer, 2006] Rolls, E. T. and Stringer, S. M. (2006). Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Computation*, 19(1):139–169.

[Roska et al., 2006] Roska, B., Molnar, A., and Werblin, F. S. (2006). Parallel processing in retinal ganglion cells: How integration of space-time patterns of excitation and inhibition form the spiking output. *J. Neurophysiol*, 95(6):3810–3822.

[Roska and Werblin, 2003] Roska, B. and Werblin, F. (2003). Rapid global shifts in natural scenes block spiking in specific ganglion cell types. *Nature Neurosci*, 6(6):600–608.

[Schindler et al., 2008] Schindler, K., Gool, L. V., and Gelder, B. (2008). Recognizing emotion expressed by body pose: a biologically inspired neural model. *Neural Networks*, 21(9):1238–1246.

[Segev et al., 1995] Segev, I., Rinzel, J., and Shepherd, G. M. (1995). *The Theoretical Foundation of Dendritic Function. Selected papers of Wilfrid Rall, with commentaries*. M.I.T. Press.

[Self et al., 2012] Self, M. W., Kooijmans, R. N., Supr, H., Lamme, V. A., and Roelfsema, P. R. (2012). Different glutamate receptors convey feedforward and recurrent processing in macaque v1. *Proc.Natl.Acad.Sci.U.S.A.*, 109(27):11031–11036.

[Shepherd, 1990] Shepherd, G. M. (1990). *The synaptic organization of the brain*. Oxford, U.K.: Oxford Univ. Press.

[Silver, 2010] Silver, R. A. (2010). Neuronal arithmetic. *Nature Reviews Neuroscience*, 11:474489.

[Simoncelli, 1999] Simoncelli, E. (1999). *Handbook of computer vision and applications, chapter 14, Bayesian multi-scale differential optical flow*. Academic Press.

[Simoncelli, 1993] Simoncelli, E. P. (1993). Distributed representation and analysis of visual motion. *Ph.D. thesis, MIT Media Laboratory, Massachusetts Institute of Technology/ Department of Electrical Engineering and Computer Science*.

[Simoncelli and Heeger, 1998] Simoncelli, E. P. and Heeger, D. (1998). A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761.

[Snowden et al., 1991] Snowden, R. J., Treue, S., Erickson, R. G., and Andersen, R. A. (1991). The response of area mt and v1 neurons to transparent motion. *The Journal of Neuroscience*, 11(9):2768–2785.

[Strout et al.,] Strout, J. J., Pantle, A., and Mills, S. L. An energy model of interframe interval effects in single-step apparent motion. *Vision Research*, 34:3223–3240.

[Sullivan and de Sa, 2004] Sullivan, T. J. and de Sa, V. R. (2004). A temporal trace and som-based model of complex cell development. *Neurocomputing*, 58(60):827 –833.

[Torje, 2002] Torje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gate pattern. *Journal of Vision*, 2(5):371–387.

[Tretiak and Pastor, 1984] Tretiak, O. and Pastor, L. (1984). Velocity estimation from image sequences with second order differential operators. *Proc. IEEE ICPR, Montreal*, pages 20–22.

[Tschechne et al., 2014a] Tschechne, S., Brosch, T., Sailer, R., Egloffstein, N., Abdul-Kreem, L. I., and Neumann, H. (2014a). On event-based motion detection and integration. *Proc. 8th Int'l Conf. on Bio-inspired Information and Communication Technologies, BICT 2014, Dec.1-3, Boston, MA, USA, ACM digital library.*

[Tschechne et al., 2014b] Tschechne, S., Sailer, R., and Neumann, H. (2014b). Bio-inspried optic flow from event-based neuromorphic sensor input. *ANNPR, Montreal, QC, Canada, Springer LNAI 8774*, pages 171–182.

[Tsui et al., 2010] Tsui, J. M. G., Hunter, J. N., Born, R. T., and Pack, C. C. (2010). The role of v1 surround suppression in mt motion integration. *Journal of Neurophysiology*, 103:3123–3138.

[Ullman, 1995] Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cereb.Cortex*, 5(1):1–11.

[Ungerleider and Mishkin, 1982] Ungerleider, L. and Mishkin, M. (1982). Two cortical visual system. *The MIT Press.*, pages 594–586.

[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):157–165.

[Vaina et al., 2001] Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., and Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceeding of The National Academy of Science 98(20)*, pages 11656–11661.

[Van Essen et al., 1981] Van Essen, D. C., Maunsell, J. H., and Bixby, J. L. (1981). The middle temporal visual area in the macaque: myeloarchitecture, connections, functional properties and topographic organization. *J. Comp. Neurol.*, 199(3):293–326.

[Van Santen and Sperling, 1984] Van Santen, J. P. H. and Sperling, G. (1984). Temporal covariance model of human motion perception. *J. Opt. Soc. Am. A*, 1(5):451–473.

[Wallach, 1976] Wallach, H. (1976). *On Perception*. Quadrangle/New York Times Book Company.

[Wallraven et al., 2003] Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with local feature: the kernel recipe. *Proc. Ninth Int. Conf. on computer vision, Nice, France*, pages 257–264.

[Wandell and Dougherty, 2006] Wandell, B. A. and Dougherty, R. F. (2006). Computational neuroimaging: maps and tracts in the human brain. *Proc. SPIE 6057*, 6057:605701–12.

[Wang et al., 2010] Wang, H., Ullah, M. M., Klaser, A., Laptev, I., and Schmid, C. (2010). Evaluation of local spatio-temporal feature for action recognition. *Proceedings of the British Machine Vision Conference BMVA Press*.

[Watson and Ahumada, 1983] Watson, A. and Ahumada, A. (1983). A look at motion in the frequency domain. *NASA Tech. Memo. (Washington, D.C.)*, 84352.

[Watson and Ahumada, 1985] Watson, A. and Ahumada, A. (1985). Model of human visual-motion sensing. *J. Opt. Soc. Am. A.*, 2(2):322–342.

[Wolf and Shashua, 2003] Wolf, L. and Shashua, A. (2003). Kernel principal angles for classification machines with application to image sequence interpretation. *Proce. Computer vision and pattern recognition*, pages I:635–I640.

[Xiao et al., 1997a] Xiao, D. K., Marcar, V., Raiguel, S., and Orban, A. G. (1997a). Selectivity of macaque mt/v5 neurons for surface orientation in depth specified by motion. *European Journal of Neuroscience*, 9(5):956–964.

[Xiao et al., 1997b] Xiao, D. K., Raiguel, S., Marcar, V., and Orban, G. A. (1997b). The spatial distribution of the antagonistic surround of mt/v5 neurons. *Cereb Cortex*, 7(7):662–77.

[Yeffet and Wolf, 2009] Yeffet, L. and Wolf, L. (2009). Local trinary pattern for human action recognition. *proceedings of the 12th international conference on computer vision*, pages 492–497.

[Yilmaz and Shah, 2008] Yilmaz, A. and Shah, M. (2008). A differential geometric approach to representing the human actions. *Computer Vision and Image Understanding*, 109(3):335–351.

[Yo and Wilson, 1992] Yo, C. and Wilson, H. (1992). Perceived direction of moving two-dimensional patterns depends on duration, contrast and eccentricity. *Vision Research.*, 32(1):135–147.

[Zeki, 1974] Zeki, S. (1974). Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *J. Physiol*, 236(3):549–573.