Towards synergistic models of motion information processing in biological and artificial vision

Naga Venkata Kartheek Medathati

UNIVERSITÉ CÔTE D'AZUR

Informatiques mathématiques Inria

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

P H D T H E S I S

submitted in partial fulfillment of the requirements for the degree of

Doctor of Science

of the Université Côte d'Azur

Specialty : Computer Science

To be defended by

Naga Venkata Kartheek Medathati

Towards synergistic models of motion information processing in biological and artificial vision

prepared at Inria Sophia Antipolis, Biovision Team

to be defended on 13th December, 2016

Jury :

Reviewers :	Ryad Benosman	-	Institut de la Vision, Paris
	Gustavo Deco	-	Universitat Pompeu Fabra, Barcelona
Examinators :	Bruno Cessac	-	Inria (Biovision), Sophia Antipolis
	Nikos Paragios	-	Centrale Supelec, Paris
	Frédéric Precioso	-	Polytech' Nice - Sophia, Sophia Antipolis
Advisors :	Pierre Kornprobst	-	Inria (Biovision), Sophia Antipolis
	Guillaume Masson	-	INT La Timone, CNRS, Marseille

Acknowledgments

This work was partially supported by EC IP project FP7-ICT-2011-8 no. 318723 (MatheMACS) and INRIA.

Contents

Abstract

Towards synergistic models of visual motion estimation in biological and artificial vision

This thesis addresses the study of the motion perception in primates. We propose that scaling up the models rooted in biological vision by taking a task centric approach would gives us further insights to probe biological vision and better constraints to design models. The first part of this thesis relates to a feedforward view of how the motion information is processed in the mammalian brains with specific focus on areas V1 and MT. Based on a standard physiological model describing the activity of motion sensitive neurons in areas V1 and MT, we propose a feedforward model for dense optical flow estimation. This feedforward V1-MT model is benchmarked with modern computer vision datasets and results form a basis to study multiple aspects of dense motion estimation. Benchmarking results demonstrated that a sharp optical flow map cannot be obtained by considering isotropic pooling and motion estimation is disrupted in regions close to object or motion boundaries. It also shows a blindspot in the modelling literature that spatial association of the extracted motion information has not been attempted or has been limited to recovering coarser attributes. In order to improve the motion estimation, we investigated the pooling by MT neurons in terms of spatial-extent and selectivity for integration as well as the decoding strategy in order to obtain a spatially dense optical flow map. We show that by incorporating a pooling strategy that is regulated by form-based cues and considering lateral propagation of the activity, the motion estimation quality is improved. Interestingly, incorporating the form based cues amounts to addition of neurons with different kinds of selectivity to the network. This raises a question, whether or not a minimal network with recurrent interactions in feature domain can exhibit different kinds of feature selectivities or we need to consider explicitly cells with different kinds of selectivity? This question relates to the second part of the thesis. We investigated this question using a ring network model under neural fields formalism with motion direction as feature space, closely mimicing MT physiological experiments. Our model produced a rich variety of results. Our results indicate that a variety of tuning behaviors found in MT area can be reproduced by a minimal network of directionally tuned cells, explicit 2D cues need not be required for motion integration, dynamical changes in the MT neuronal tuning reported in the literature can be explained through feature domain recurrent interactions and also open the door for accounting transparency by challenging the high inhibition regimes considered by many models in the literature for motion integration. To conclude, we re-emphasize on task-centric modelling approaches and several directions for interfacing studies in biological and computer vision.

Chapter 1

Introduction

"There are things known and there are things unknown, and in between are the doors (of perception)."

- Aldous Huxley, *The Doors of Perception*

What is visual perception?

We are in a constant need to be aware of the environment around us for our survival and well being. We need to know where we are, what things are around us and where the things around us are moving. In order to be aware of the environment, we rely upon early sensory mechanisms that capture changes in the environmental energy in various forms such as patterns of illumination (visual), patterns of vibrations (auditory), patterns of pressure (somatosensory) etc. While vision is one of dominant senses, the sensory information that is being captured in the form of illumination patterns is often incomplete, noisy and ambiguous. Due to these difficulties, attributes of interest such as objects that are present or their motion information are not directly available at the sensing level. Consider the scenes depicted in Fig. 1.1, it is a non-trivial challenge to detect what objects/animals are there and where they are from an array of intensity samples of the scene. Visual perception or in brief vision refers to the active process of identification, interpretation and organization of the visual information overcoming the inherent sensory ambiguities in order to become aware of the environment.

Figure 1.1: Illustrating few scenes where our visual system needs to overcome ambiguities in order to extract relevant information. We need to carefully group parts of the image with distinct visual appearance to identify an object at the same time we need to separate parts which are close in appearance but belonging to different objects.

Understanding how the visual system performs these functions is a fascinating question and has been investigated for a long time by several different disci-

plines [Palmer 1999]. One might even start asking a much more fundamental question, what does it mean to understand vision? David Marr's [Marr 1983] suggestion to this question is that any complex information processing system such as visual system needs to be understood at three distinct levels, computational, algorithmic and implementational. Computational level relates to understanding the computational problem that the system is handling, algorithmic level relates to identification of the strategy that the system adapts to solve the computational problem and implementation relates to characterization of how the algorithm is physically executed.

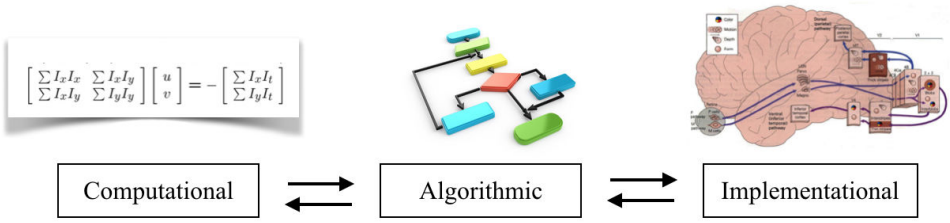
1.1 The Goal

The goal of this thesis is to understand low level visual motion perception by developing computational models that describe early visual motion estimation. It would mean to identify potential algorithmic strategies that could be adapted by the system as well as plausible neuronal implementation. In short, characterization of representation and information processing carried out by areas specialized in low level motion processing within the primate visual system.

1.2 The Challenges

There are many challenges in developing computational models that describe motion perception.

- Interdependence of different levels: Even though Marr's [Marr 1982] three levels of understanding framework provides a powerful and intuitive way to explore the visual system, these three levels are not independent.

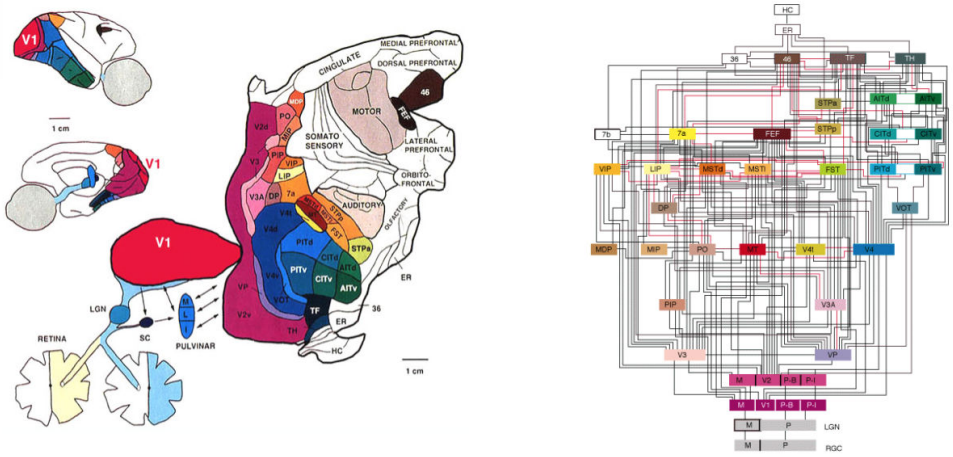


Three levels of understanding

What is the right computational problem? Identification of the computational problem is a primary step in order to be able to probe the visual system in terms of the algorithmic strategies and structural implementation. In the context of motion processing, a lot of progress has been made on this front [Fennema 1979, Bradley 2008] (see Chapter.2 for a detailed discussion). Going from the level of computational problem to identification of the algorithmic strategy is tricky. The visual system might not be relying on a single

strategy and could be changing the algorithmic approach in a context dependant manner. The computational problems also manifest in different forms with respect to the algorithmic strategy that is being adapted. Even after identification of a plausible algorithmic strategy, relating it to the neuronal activity would depend largely upon on the assumed representation. Thus it is very difficult to come up with generic models that could describe the motion perception in complex naturalistic scenarios.

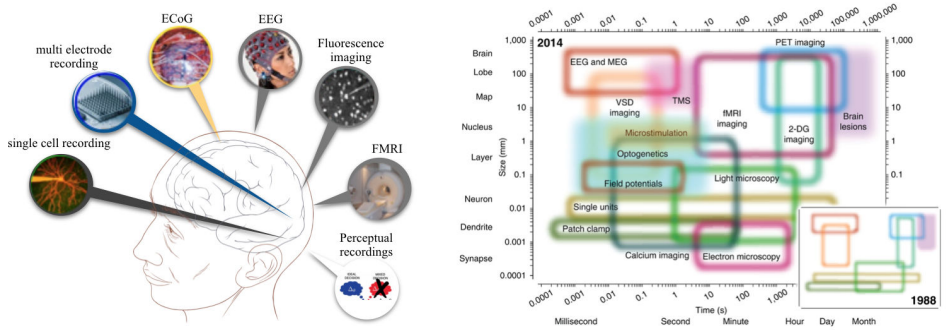
- Complexity of the network: The primate visual system is phenomenally complex [Felleman 1991]. It occupies around 52% of mammalian cortex, has several recurrently interconnected areas and is capable of solving several ill-posed problems such as segmentation, recognition, motion and depth estimation within a matter of few milliseconds.



Flat map of the macaque monkey brain and hierarchy of the visual areas. Figures taken from [Felleman 1991].

What are the effective ways of abstraction to study such a complicated network? To simplify the problem, traditionally the network has been divided into "what pathway" and "where pathway", each comprising of few functionally specialized areas. Even this kind of coarse abstraction is challenged by recent evidence suggesting significant amount of interactions between areas assigned to two different pathways. Even within each area, there are diverse cell types with varying selectivity. The selectivity of the cells has also been found to change dynamically and also depended on the type stimuli used to probe them. Under these observations, selection of an appropriate minimal network representative of the areas involved is challenging. Even after selecting few representative areas, one needs to think about the modes of interaction between the network elements that need to be taken into account. Consideration of feedforward only or inclusion of recurrent and lateral interactions is a difficult choice. This question would also translate into selection of appropriate representation.

- Limited observation windows: The functioning of the visual system is probed using a variety of techniques such as single/multi electrode recordings, functional imaging using techniques such as VSD imaging, functional MRI, behavioural responses such as eye tracking or perceptual reports. Each of these techniques yield observations related to the activity of the brain at different spatio-temporal scales and often do not provide detailed information for characterizing representation/information processing strategies. Another aspect is that not all techniques are available to probe all the areas of interest as some of them might be difficult to reach.



Techniques used to probe brain function and their effectives in terms spatio-temporal resolution of observations [Sejnowski 2014].

In other words, in order to understand how the visual system solves the ambiguities in the sensory information, one needs to overcome the observational ambiguities introduced by the limitations of the available techniques to probe the functioning of the brain.

1.3 Methodology

Given the limitations of the experimental techniques, theoretical analysis and computational modeling play an important role in linking different observations. Theoretical models are helpful in constructing compact representations and building bridges across different levels of description. Theoretical models of perception have been classified into three different categories [Dayan 2001], depending on which of the what, how or why questions they answer. Descriptive models suggest algorithmic rules that can potentially explain large amount of experimental data. In the context of visual motion estimation one can give example of rules such as Intersection of Constraints [Fennema 1979] or Harmonic Vector averaging [Johnston 2013], which describe perceived motion direction when there are multiple components of motion present in the stimuli. Mechanistic models address the question of how systems operate in a bottom-up manner and often attempt to descibe how descriptive models could be implemented by the neural networks establishing a structure/function relationship. In the context of motion estimation examples of such models could

be the pattern cell model by Simoncelli and Heeger [Simoncelli 1998]. Interpretive models try to explain the behaviour of a system with a top-down approach focusing on the functional role of a phenomenon or why we tend to observe the phenomenon. These models could be explaining why a particular approach has been adapted by the visual system. In the context of motion estimation, example of such model could be Bayesian estimation [Weiss 1998] where the visual system is continually trying to minimize the ambiguity in the observations.

Eventhough theoretical models have the potential to establish structure/function relationship, there is currently a lot of fragmentation in these models. The models are rooted heavily on explaining a particular experiment, for example, empirical rules such as Intersection of Contraints describe the eventual psychophysical percept that is being reported whilst ignoring the representational questions. One cannot decipher whether the visual system computes a dense flow field which is uniform for the stimuli or selects regions within the stimuli which are not ambiguous to arrive at the percept. If we consider a mechanistic model such as Simoncelli-Heeger [Simoncelli 1998] model, the model explains how a particular MT cell could exhibit pattern motion selectivity but does not elaborate on how population of MT cells could encode/represent different scenarios such as transparency, motion boundaries etc., which need to be dealt with in order to interpret motion information. Overall, within the context of each of these models, the task of motion estimation is reduced to several sub-aspects or "read-outs" that a particular experimental technique could provide, thus ignoring key representational questions and overall effectiveness of the proposed strategies in terms of recovering motion information.

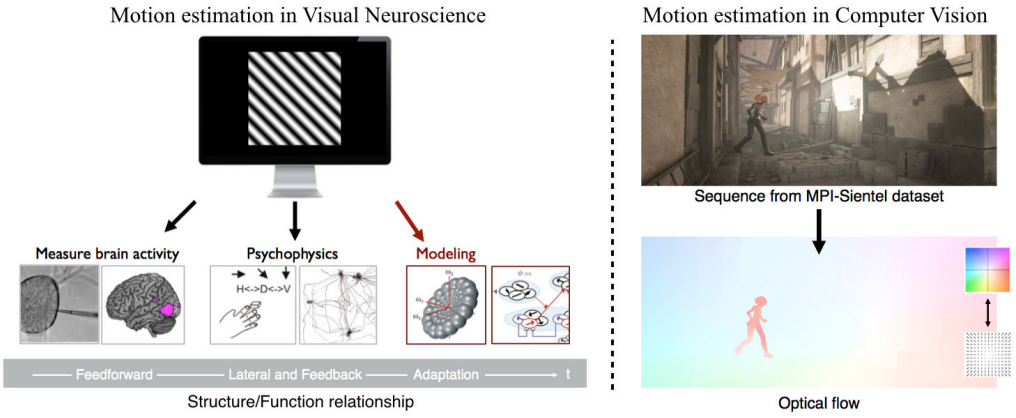


Figure 1.2: Illustrating approaches taken by studies in Visual neuroscience and Computer vision towards motion estimation.

There is a complementary approach that has been taken in the field of computer vision, where researchers have heavily focussed on the view point of recovering dense motion vector map for all the elements within the scene. This focus has lead to the development of public domain datasets resembling complex naturalistic scenarios

and benchmarking practices where efficacy of the models/algorithms could be tested on standard criteria. Fig. 1.2 illustrates the current scenario in tacking the motion estimation problem from the view point of computational neuroscience and computer vision.

Historically, the stimuli that are being explored in neurosciences were carefully tailored to test some of the computational questions that a visual system might encounter in natural scenes [Albright 1995]. Consider the naturalistic scene depicted in Fig. 1.3, where the visual system has to detect motion and has to do non-trivial interpolation and segmentation before the scene could be decomposed into meaningful motion components. Synthetic stimuli such as plaids and moving gratings allowed the neuroscientists to ask critical questions about the rules that visual system could rely to integrate or segment motion components. However, the models ignored the aspect of dense recovery and did not make a rigorous attempt to verify the scalability of the rules that were derived using the synthetic stimuli and thus could be missing out on important constraints that need to be considered. Part of the reason could have been the lack of appropriate ground truth in naturalistic scenarios. This bottle neck has long been overcome by Computer vision.

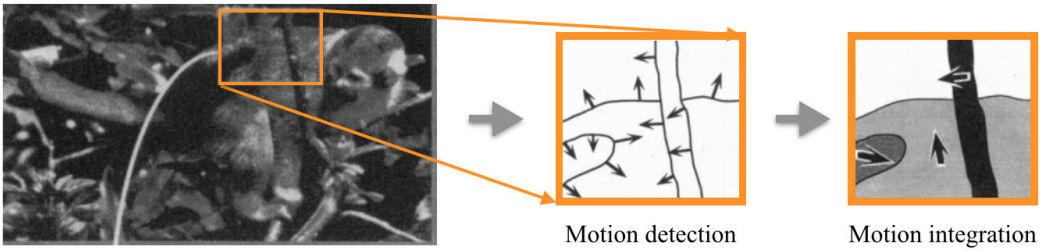


Figure 1.3: Illustrating a natural scene where the visual system has to disambiguate motion signals from multiple sources by integrating and segmenting appropriately. Figure adapted from [Albright 1995].

In this thesis we take a task-centric view for the motion estimation problem, the idea is to scale up models rooted in visual neuroscience and derive a dense flow optical flow map using them going beyond the stage of "read-outs" or synthetic stimuli they were originally designed for. This would rigorously test the assumptions made by models and could provide new insights into probing the visual system itself. We also examine what could be implications of taking task-centric view to interface studies in neurosciences and computer vision in general.

1.4 Organization and main contributions

The thesis is organized into four parts.

PART I is a review of the studies in low level motion estimation. We begin by summarizing the computational problems that visual system encounters towards

dense motion estimation, we present relevant findings from the experimental investigations at both psychophysical and physiological levels. We briefly present key mathematical models that describe the experimental observations related to motion detection and integration. We discuss open questions that existing models do not address.

PART II is centred on scaling up models rooted in biology for motion estimation in real world scenarios using feedforward filtering framework. We describe a minimal feedforward model representative of pattern motion selective cells in MT area and evaluate the model performance using standard computer vision dataset. A form based pooling mechanism has been developed and it is demonstrated that it improves the performance of the model. A regression based decoder has also been studied to exploit machine learning techniques to improve the flow estimation. We discuss a key question, whether we need to incorporate additional cell types or the network effects in the form or recurrent interactions have to be better captured?

PART III we explore the role of interplay between recurrent interactions and driving input on tuning properties of the cell within a network using neural fields formalism. Using bifurcation theory and a structured input representative of standard stimuli used in physiological and psychophysical experiments we characterize the behaviour of the model under various connectivity regimes. The numerical results are used to explain various kinds of temporal dynamics and behavioural switches observed in MT tuning properties.

PART IV We present task-centric summary of models in neuroscience by considering additional tasks, namely sensing and segmentation, followed by perspectives on interfacing biological and computer vision and a publication list resulting out of this work.

Part I

Studies in visual motion processing

Chapter 2

Visual Motion Estimation

"The level of computational theory is fundamental to understanding vision. From an information processing point of view, computations that underlie perception depend more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented."

- David Marr, *Vision*

"Computational theory enables us to formulate a number of empirical questions that would not otherwise arise, and it opens the way for a rational investigation of the phenomenon rather than the confused cataloguing of its phenomenology."

- S. Ullman, *The interpretation of structure from motion*

We live in a dynamical world, in which detection, intepretation and organization of the changes in the environment are essential for survival. Vision motion information serves several functional needs in this regard [Nakayama 1984]. For example, it is used to detect an approaching predator or to track and hunt a prey, for navigation, to understand ego-motion etc. However, the motion information is not readily available to the visual sensing devices, in either biological or artificial systems. It has to be estimated from the spatio-temporal patterns of light that is projected onto the two dimensional sensor (retina/camera) from the moving world. The problem of estimating the displacement of the scene elements, pixels or objects in the image plane is referred to as *optical flow estimation* and is a non-trivial problem with several computational challenges.

The computational difficulty is also evident from the fact that biological visual systems, ranging from small insects to complex ones like primates, devote considerable amount of neural resources to motion processing. Owing to the scientific curiosity and application potential, this problem has been well studied in to both artificial and biological vision. Although, it is clear that processing constraints in biological systems are different from those in an artificial vision system, they share the goal of extracting motion information from a very similar input of temporally varying intensity patterns. Given the evidence that computations analogous to optical flow estimation are performed by low-level sensory processing systems in biological vision, we consider understanding flow computation as a first step in understanding motion perception. In this chapter, we begin by summarizing the computational

challenges that underlie visual motion estimation with the conviction that identification of the fundamental computational challenges would give us a way not only to provide an organized presentation of the literature but also a key to probe biological vision solutions and to test the efficacy of the proposed models.

2.1 Computational challenges in motion estimation

We can appreciate the difficulties in optical flow estimation by analysing few cases: consider a simple scene containing few dots moving diagonally upward as shown in Fig. 2.1. The task of optical flow estimation would be to assign displacement vectors to each of the dot within the scene, thus one needs to be able to match the dots in the frame captured at time ($t1$) to the one captured after at time ($t1 + \delta t$). Assuming rigid motion, this assignment is achievable only as long as the dots are distinct, thanks to either brightness or color as shown in Fig. 2.1.a. But, if the dots have similar appearance, one can not reliably estimate the displacement vectors without taking additional constraints into consideration as the matching is ambiguous as shown in Fig. 2.1.b.

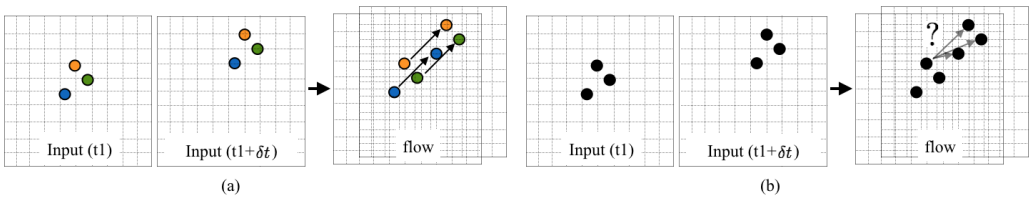


Figure 2.1: A scene illustrating the ambiguity in optical flow estimation. In (a), the motion of the dots is uniquely identifiable because correspondence could be established between the pixels across the temporal samples. In (b), a variety of motions could result in the configurations observed making the estimation ambiguous unless additional contextual information is available to estimate the flow.

Considering the surrounding of a pixel is helpful to resolve ambiguities in motion estimation as long as there are distinct patterns of brightness changes in its vicinity. For example, in Fig. 2.2.a, unique motion vectors could be estimated by considering that a group of dots within a small window have similar motion. Even after considering local context of a pixel, absence of 2D pattern features still renders the motion estimation problem as ill-posed and gives raise to two different computational problems referred to as *the blank wall problem* and *the aperture problem*. The scenario in which there is no brightness change is referred to as the blank wall problem: in this case the motion is not observable for a local detector as illustrated in Fig. 2.2.b. The scenario in which there is brightness changes only along one direction is referred to as aperture problem: in this case the velocity estimate along the direction of the contrast ambiguous as illustrated in Fig. 2.2.c. Note that the aperture problem is a much more generic term and is used widely in other scenarios where the observation window is limited and thus renders the available information ambiguous.

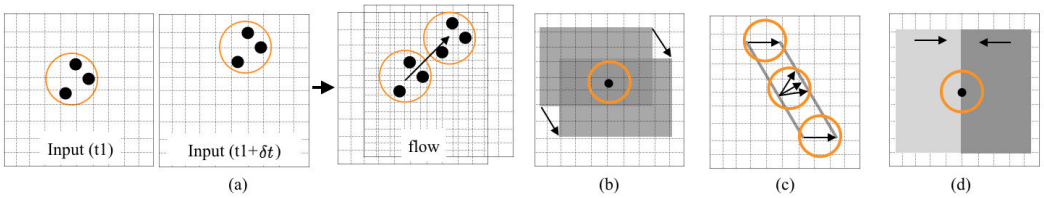


Figure 2.2: Illustrating scenarios considering local contextual information to resolve ambiguities in optical flow. In (a), the motion of the dots is uniquely identifiable by assuming that dots within a small neighbourhood (in orange) to have similar motion. In (b), considering local neighbourhood does not help in the absence of local contrast, known as *the blank wall problem*. In (c), velocity is resolvable only along the direction orthogonal to contrast. In (d), assuming uniform motion in a local neighbourhood would lead to erroneous estimates due to presence of multiple motion components.

The third aspect pertains to the selection of the local context itself. The selection problem arises when there are instances in the scene where multiple components of the motion are present within the neighbourhood of the pixel. Such instances can be observed at occlusion boundaries, where the object is in motion and background is stationary or motion boundaries where there are multiple motion surfaces in the vicinity as illustrated in Fig. 2.2.d or an extreme case of transparency, where multiple motion surfaces are present/perceived at the same location. In all of these scenarios, selection of the patterns to establish a context and representation of motion surfaces are difficult problems to address.

2.2 Investigations into biological vision

How does the brain solve these problems for motion estimation? What kind of algorithmic approaches does it take? What are the neural substrates implementing the approach? These questions have inspired a variety of studies using rich variety of techniques and stimuli. Here, we briefly overview psychophysical and physiological explorations of these questions. Psychophysics is particularly interesting as it is non-invasive and has great potential to give insights into the representational and information processing strategies. Physiology gives us direct access to probe neural computation.

2.2.1 Psychophysics

Early investigations into the computational mechanims underlying motion estimation came in the form of behavioral experiments involving Chlorophanus beetle by Hassentein and Reichardt [Reichardt 1961]. The beetle was glued to a stick and mounted on a y-globe. It was presented with moving stimuli and was tested for its behavioural responses such as direction in which it turns in response to the stimuli, see Fig. 2.3 for an illustration. The observations lead to the formulation of a basic

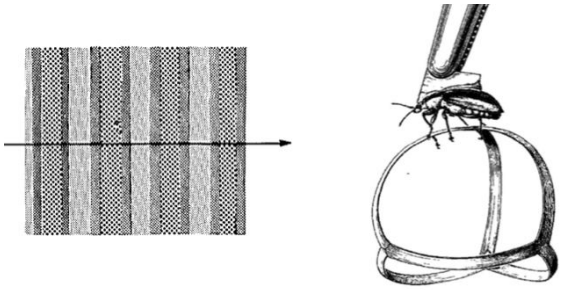


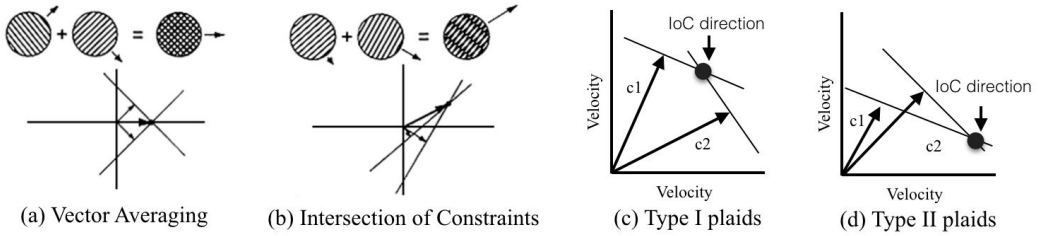
Figure 2.3: Illustration of the early behavioral experiments on insects leading to the formulation of correlation type detectors. Figure adapted from [Poggio 2011].

correlation type detector, where the inputs from two spatially separate samples are compared against each other to estimate motion. The fundamental structure of the correlation-type detectors has been modified further to account for the behavioral observations from human perceptual experiments involving moving sinusoidal gratings of varying contrasts as input [Van Santen 1984], where the subjects were requested to report the direction of motion that is being perceived. Currently, it is a dominant view that early motion detectors are most sensitive to motion orthonogal to the local image constrast. It would imply that the visual system could be generating several 1D motion cues, as the detectors are not very sensitive to motion information along the contrast orientation. However, the motion information is ambiguous when only 1D cues are considered, which has long been known to the experimentalists [Wallach 1935]. The system has to combine multiple cues that are available in the stimuli to eliminate ambiguity and reliably estimate motion.

2.2.1.1 Phenomenological rules describing 1D/2D cue combination

How does visual system disambiguate motion? There are two aspects to this question, how multiple 1D cues available locally could be combined and how 2D motion signals are integrated across space and time to yield a percept? The first question has been studied extensively using spatially overlapping moving gratings, known as plaids, in various configurations [Adelson 1982, Gorea 1991, Yo 1992]. Plaid stimuli are ideal to study this question as they are analogous to natural conditions of oriented contours moving in front of each other. Interestingly, the behavioural response of the visual system to plaid patterns is not-so straight forward to interpret and depends on the parameters of the stimuli.

Two possible rules have been proposed for determining perceived motion direction in terms of the velocities of the constituent components in plaid patterns, the system could take a simple vector average (VA) or follow a more complex approach in the form of intersection of constraints (IoC) as illustrated in Fig. 2.4. The plaid stimuli themselves are classified into two categories based on the position of the direction vector predicted by IoC rule with respect to components. In Type I plaids,



(a) Vector Averaging (b) Intersection of Constraints (c) Type I plaids (d) Type II plaids

Figure 2.4: Figure illustrating plaid patterns and associated rules that describe component integration.

IoC direction is located in between the component velocity vectors. In Type II plaids, it falls on the same side. This has been illustrated in Fig. 2.4 (c and d). [Wilson 1992] claim that there is a qualitative difference in the appearance of the Type I and Type II plaids: Type I plaids seem to be "rigid" whereas Type II plaids apear to be "fluid like (blobs)". Although the two types of plaids differ perceptually, the main reason for the distinction is that in Type II plaids direction predicted by VA is significantly different from the one predicted by IoC and the perceived direction tends to be closed to the IoC direction. In unikinetic plaids, an extreme case, where one of the grating pattern is stationary, the IoC rule is not valid, as there is only one component in motion. However, the percept could be explained by the motion of the 2D features (blobs at intersections) and stands as an evidence that both a combination of 1D velocity estimation mechanisms and pattern velocity (blobs) information are being used by the visual system simultaneously [Gorea 1991]. This re-introduces the debate whether 2D motion cues are always decomposed into 1D constituents in the early stages of detection? An instance of plausible change in the algorithmic approach based on the context.

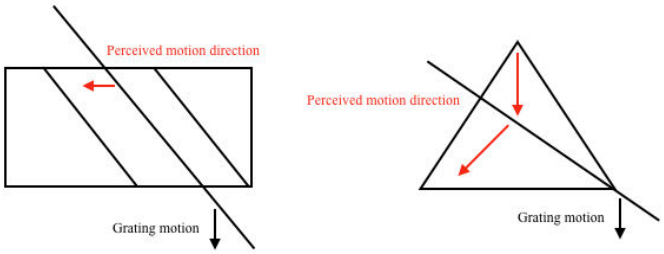


Figure 2.5: Figure illustrating perceived changes in direction of motion of the grating due to the change in 2D cue information influenced by the shape of the aperture. Figure adapted from [Wallach 1935]

Various forms of interactions among 1D and 2D cues can be readily seen if we look into the ingenious experiments performed by Hans Wallach [Wallach 1935, Wuerger 1996], where drifting lines and gratings under different apertures were presented to the observers. It has been reported that perceived motion direction of

the line was along the orientation of the edge of the aperture it intersected. If the two aperture edges that intersected simultaneously had different orientations, then the percept was reported along an intermediate direction. Sharp transitions in the percepts were also reported if the intersection angles changed due to the the shape of the aperture, these are illustrated in Fig. 2.5.

The selection or integration choice has also been found to be influenced by segmentation cues available with-in the images. For instance, we know from physics that when two transparent surfaces are physically overlaid, the resulting intensity that could be observed is neighter too bright nor too dark. This kind of prior information could be taken into account by visual system in grouping the available cues. [Stoner 1990] tested this using plaid patterns by modulating the luminance at the intersections. They found that if the luminance levels at the intersections were within a range to support two transparent surfaces being overlaid, transparent motion is perceived. Breaking away from that luminance range either by setting intersections to be too dark or two white increased the likelihood of a coherent percept. The stimuli are illustrated in Fig. 2.6.

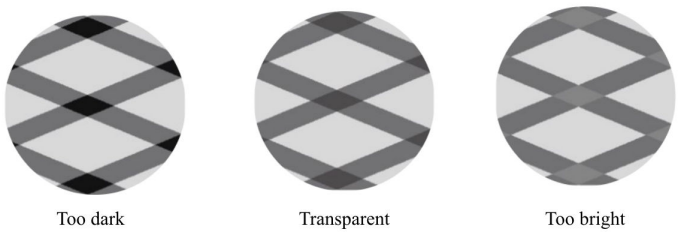


Figure 2.6: Figure illustrating modulations of the intensity at intersections of the plaid patterns. If the intersections are too dark or too light, the percept is found be dominantly coherent and in between the percept is found to be dominantly transparent. Figure adapted from [Stoner 1990]

Similarly, the visual system seems to selectively consider available 2D cues generated, at times even discarding some of them. This can be readily observe in case of the chop-sticks illusion [Anstis 1990], as illustrated in Fig. 2.7. In the configuration where the line-endings are visible, the 2D cues generated at the intersection of the moving bars are discarded and the bars are perceived to be sliding over each other. In the configuration were line endings are occluded by an external surface, the 2D cues generated at these intersections are discarded and the bars are perceived to be moving together along the direction of motion of their intersection.

In short, the motion percept seems to be a result of interactions among different kinds of cues that are available in the stimuli. If several different 1D cues are available locally, then cues are integrated along the feature space (motion) and in the regions where local cues are inadequate, 2D cues from distant locations are being relied upon. In this process, contextually the 2D cues are combined, discarded or selected based on form based cues.

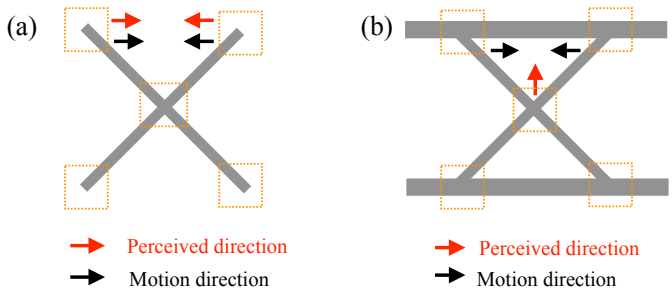


Figure 2.7: Illustrating contextual selection of 2D motion cues. (a) Percept is dominated by 2D cues at the end of the bars. (b) Percept is dominated by 2D cues generated due to intersection of the bars. The available 2D cues are highlighted using dotted orange boxes.

2.2.1.2 Dynamical changes related to motion integration

The analysis of the reported percepts could shed a light on the eventual solution adopted by the visual system, it does not shed light on how the visual system arrives at the solution itself. In order to gain deeper insights, the temporal evolution of perceptual capacities has been exploited by psychophysicists. Perceptual computations are better understood by measuring reaction times, limiting viewing times, or using clever tricks such as masking to interrupt perceptual processes at different times.

2.2.1.3 Early dynamics in solving aperture problem

[Castet 1993] probed the estimated direction of motion of moving bars of varying lengths when they are presented for very short intervals of about 170ms and found that accuracy decreases as the length of the bar increases. [Wallace 2005, Born 2006] recorded the behavioural responses in the form of eye tracking when subjects were asked to follow the center of a moving bar in case of both humans and trained monkeys. These experiment revealed a striking trend in terms of accuracy of the motion direction estimation, with respect to time. Initially the ocular flow response followed the direction orthogonal to the tilt of the bar and got slowly adjusted to the right direction with gradually decreasing angular error. This is very striking as it hints at the possibility of visual system solving the aperture problem by propagation of non-ambiguous cues across stimuli rather than just selecting informative areas in the stimuli and ignoring the task of dense field estimation. These observations are illustrated in Fig 2.8.

In the case of translating bars, there are only two kinds of cues, ambiguous 1D cues at the center and unambiguous 2D cues at the edges. Translating diamonds are bit more interesting stimuli, bars present with in the diamond have different slants, thus the 1D directional estimates from the ambiguous parts are different. We thus have a scenario of two contrasting 1D cues, along with corners which provide unambiguous 2D motion estimates. In this case, the initial ocular responses were

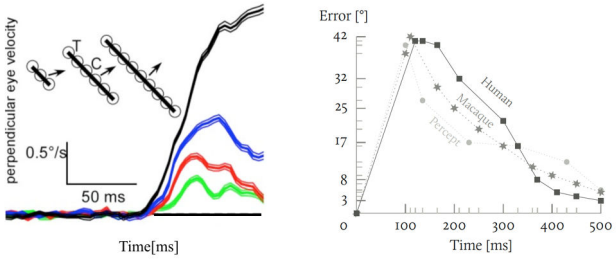


Figure 2.8: Illustrating early dynamics of motion estimation reflected in ocular response to moving bars, adapted from [Tlapale 2011a].

found to follow a vector averaging rule and then later on, have a course correction towards true direction of the object [Wallace 2005] better predicted by IoC rule. This observation once again re-emphasizes the fact that local ambiguity is resolved by considering all the available cues.

2.2.1.4 Slow changes in the percept over prolonged stimulation

Setting aside early dynamics, at a slower time scale, prolonged stimulation (10s of seconds) using certain stimuli was found to elicit multiple percepts over time, this is referred to as perceptual multistability. When, it comes to motion perception, two important stimuli could be discussed. One being a classical barber pole and other being plaid patterns viewed under circular aperture. When a drifting grating patterns are observed through a symmetric square aperture, under prolonged exposure, horizontal, vertical or motion orthogonal to grating orientation are perceived [Wallach 1935] as illustrated in Fig. 2.9.

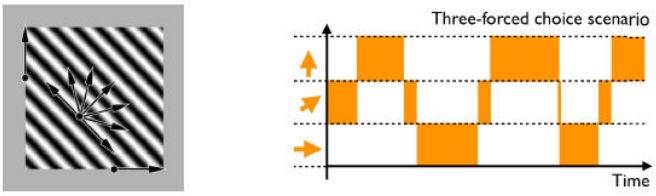
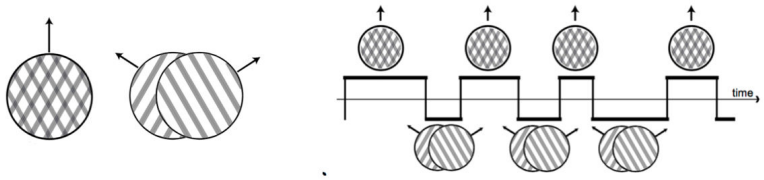


Figure 2.9: Illustrating a square barberpole stimulus and reported perceptual switches overtime.

In the case of plaids, the visual system could either group the two component motions as a single pattern or could treat the two moving components as separate ones and consider the motion surfaces to be transparent. Experimentally, it has been found that coherent pattern motion along the direction predicted by IoC is perceived initially. Upon prolonged stimulation, the percept was not only found to change to a transparent one but also to switch between these two states of coherency and transparency periodically as illustrated in Fig. 2.10.



(a) Plausible percepts (b) Perceptual switches over time

Figure 2.10: Illustrating possible percepts for a plaid stimuli and perceptual switches overtime.

2.2.2 Physiology

In this section, we present a brief overview of vast amounts of literature that concerns neural correlates of the motion estimation with focus on areas V1 and MT, as they have long have been associated with low-level motion percepts. The early discoveries of neural correlates to motion detection came from the seminal experiments by [Hubel 1965, Hubel 1968], who discovered cells in Cat's visual areas 18 and 19 which responded to motion of spots and oriented contours followed by similar observations in the visual area V1 in striate cortex of the monkey. A brief illustration of the observation is presented in Fig. 2.11. The cells respond by firing vigorously when an oriented bar moves with in small region (usually perpendicular to orientation) but are silent when the motion direction is in the opposite direction.

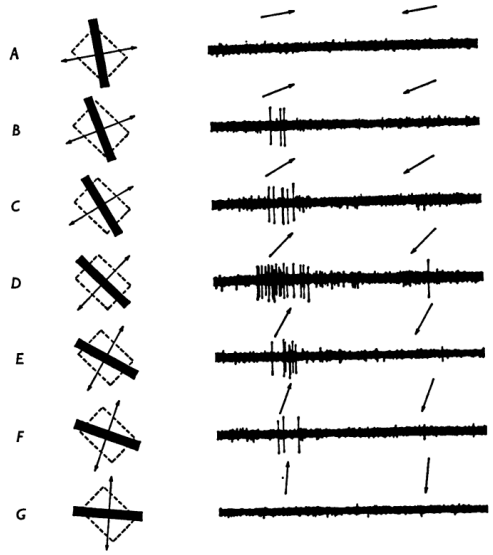


Figure 2.11: Illustrating the responses of V1 direction sensitive cells when stimulated with oriented bars moving in different directions.Figure adapted from [Hubel 1968].

These observation lead to the early models of the motion detection by [Movshon 1978] and later were found to be analogous the models describing

motion perception in case of gratings and sinusoids by [Adelson 1985]. Thus, linking percepts and physiological responses in cases of simple stimuli such as moving gratings and leading to the popularity of spatio-temporal energy models. These early discoveries motivated further experiments with increasingly complex stimuli. A natural and popular extention to the stimuli such as moving bars and gratings is plaid patterns: physiological experiments using plaids revealed a variety of responses in area MT [Movshon 1985].

2.2.2.1 Motion tuning properties of MT neurons

The middle temporal visual area (MT or V5) is rich with cells which are responsive to moving stimuli and are often selective to the direction of motion independent of orientation of the underlying contrast. This is different from cells in V1, which respond well when the direction of motion is orthogonal to the contrast. Thus, MT is believed to the region in the brain where aperture problem is solved. But, understanding the response selectivity of MT is not so straight forward. The cells are found to change their response selectivity dynamically and also with respect to the stimuli used to probe them. Here, we brief few observations of interest.

2.2.2.2 Prominent tuning characteristics

Pattern and Component cells Traditionally preferred direction of MT neurons is tested using moving gratings and then the cells are further probed using stimuli such as plaids [Movshon 1985] or overlapping RDKs [Snowden 1991]. However, the cells are classified into two different classes based on their tuning properties to plaids (comprised of overlapping grating stimuli). The cells which exhibit a unimodal response in the direction domain to the plaid patterns are referred to as *pattern cells* as they are tuned to direction of motion of the pattern as whole and are believed to be solving the aperture problem. Cells which respond with bimodal lobes to plaids are referred to as *component cells* and are believed to play a role in perception of transparency [Snowden 1991].

Side Bias or Component selectivity For a long time, the above two types of tuning were thought to be representative of component motion or pattern motions in the driving stimuli. However, recent investigations by [Xiao 2015] have shown one more type of tuning where cells preferred one of the components instead of pattern selectivity or being reponsive to either of the components. Fig. 2.13 illustrates the different kinds of selectivities obtained by [Xiao 2015] by averaging the firing rate responses over a period of 1000 ms when cells are stimulated with overlapping RDKs with two different directions of motion.

To summarize, at least, three distinct types of tuning behaviour are observed when MT cells are probed with stimuli comprising of overlapping motion components. The cells could merge the constituent components, respond to either of the components or select one of the components, which is analogous to what has been

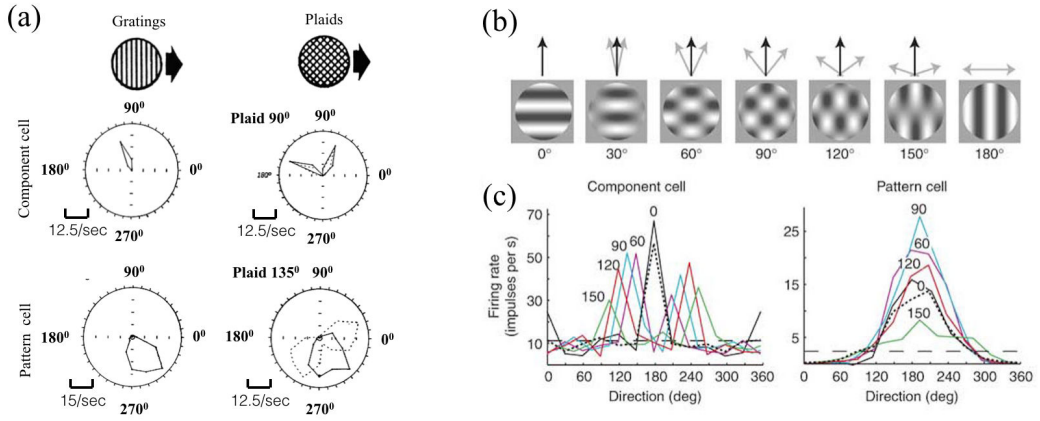


Figure 2.12: Illustrating traditional classification of MT cells based on the qualitative tuning responses to gratings and plaid stimuli. (a) Experimental data showing the tuning response of component cell and pattern cells as a polar plot representing the firing rate to grating and plaid stimuli, adapted from [Movshon 1985]. (b) Illustrating various plaid patterns formed by varying angles between the constitutent components, (c) Illustrating responses of sampled pattern and component cells, component cells exhibiting a bimodal tuning curve with respect to motion direction and pattern cells exhibiting a single peak. (b) and (c) adapted from [Rust 2006].

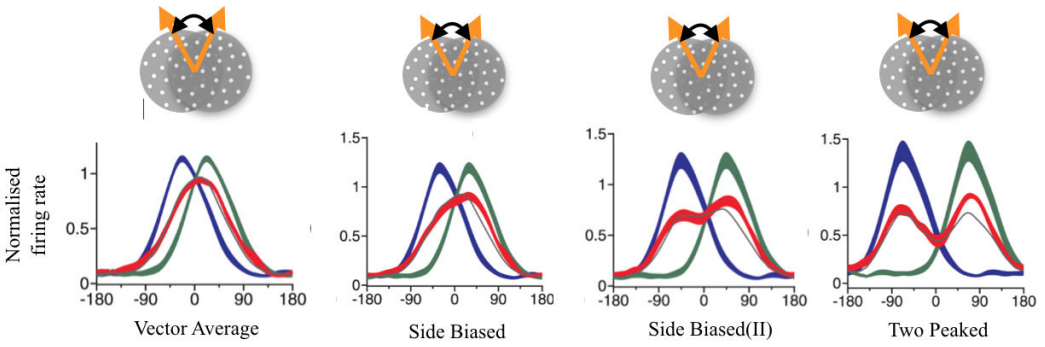


Figure 2.13: Illustrating different kinds of tuning curves exhibited by MT direction selective cells when stimulated with a set of overlapping RDKs. In blue and green are the responses of the cell when only one of the components of the motion is present and in red are the response of the cells to overlapping random dots. Figure adapted from [Xiao 2015].

also observed psychophysically. How each of these tuning behaviours is relevant in solving the aperture problem or representing pattern versus transparent percepts is an open question.

2.2.2.3 Temporal dynamics

MT cells are found to exhibit several different kinds of dynamical changes in the tuning behavior: their response selectivities were found to emerge temporally and their tuning properties were shown to change depending on the input type and context [Pack 2008]. [Pack 2001] using an array of moving bars as stimuli have demonstrated that early responses of MT neurons are dependant upon the contrast orientation while late responses show true motion direction selectivity independent of the orientation of the bar as illustrated in Fig. 2.14.

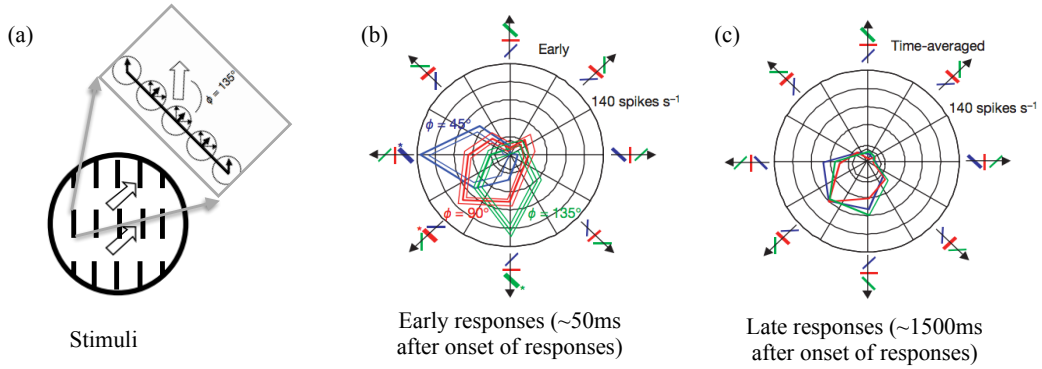


Figure 2.14: Figure illustrating temporal emergence of the tuning behavior with oriented bars. (a) An array of moving bars stimuli used to probe MT tuning, (b) Early responses of the MT neurons demonstrating that neuron has a preference to left-oblique bar either moving downwards or leftwards, showing dependence on contrast orientation, (c) Late responses demonstrating the motion direction preference of the neuron independent of the contrast orientation. Figure adapted from [Pack 2001].

[Pack 2004] demonstrated that MT firing rate or response selectivity changes slowly in the presence of an aperture. It has been shown that the aperture has no impact on the neuronal selectivity for the first 40-60 ms and if one considers a longer time window of about 200-1000 ms, the neuron seems to fire for any motion direction that contains a motion component along the elongated edge of the aperture. The Fig. 2.15 shows the temporal changes in the tuning of MT cells when they were recorded with moving gratings presented behind an aperture.

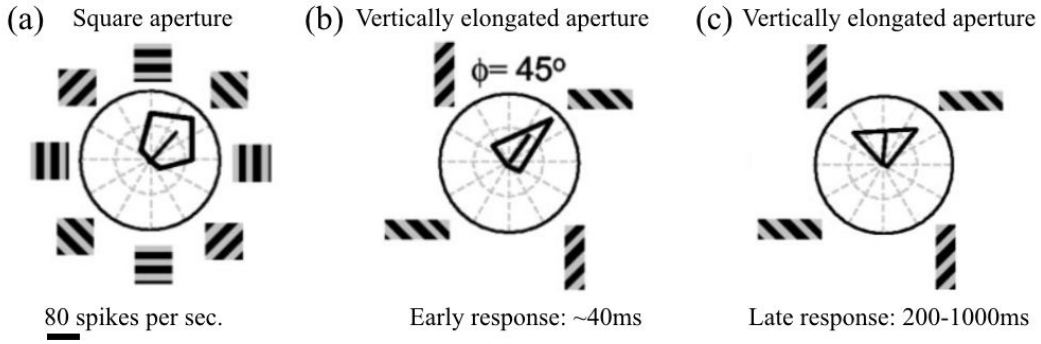


Figure 2.15: The aperture shape influences the late response tuning characterisitics of the MT neurons. (a) The prefered direction of motion for the MT cell when the grating is presented under a square aperture. (b) In the initial time period of around 40ms, the MT neuron still responds to the prefered direction $\phi = 45^0$ even when the grating is presented under a vertically elongated aperture showing little impact of the aperture shape. (c) Late response, over a period of 200-1000ms the cell is shown to be reponding to any motion direction which has a component of motion along the elongated side of the aperture. Figure adapted from [Pack 2004].

While the previous two instances of the temporal changes in the tuning could be attributed to the spatial integration of the cues for the aperture, [Smith 2005] have demonstrated that even the characteristic tuning behaviours exhibited by the MT neurons such as pattern and component selectivity emerge over a time of 60-80 ms. Considering the z-score, which computes the similarity of the tuning curve with either pattern like response or a component like response, they elucidated that component cells respond faster than pattern cells by about 6 ms and both pattern selectivity and component selectivity of the cells emerge gradually as shown in Fig.2.16.

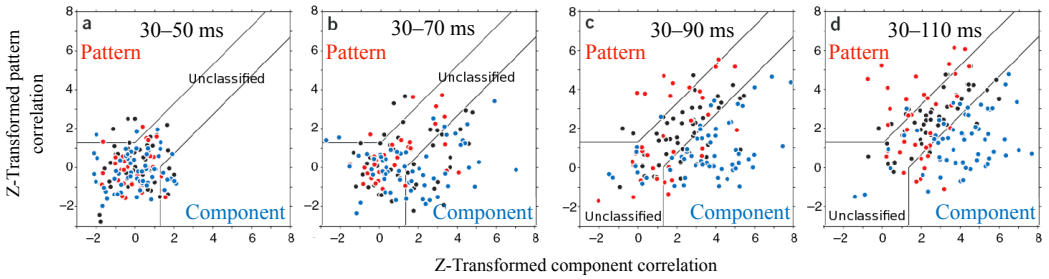


Figure 2.16: Illustrating the temporal emergence of the response selectivities of pattern and component cells based on z-score computed over different time windows, figure adapted from [Smith 2005].

Similarly, [Xiao 2015] have demonstrated the temporal evolution of tuning using overlapping RDKs as illustrated in Fig.2.17, which by far is the most detailed

presentation of the dynamically emergent behaviour of the MT direction selectivity. Initially the responses appear to be biased towards a VA direction and later on could emerge either as a side-biased or component selective.

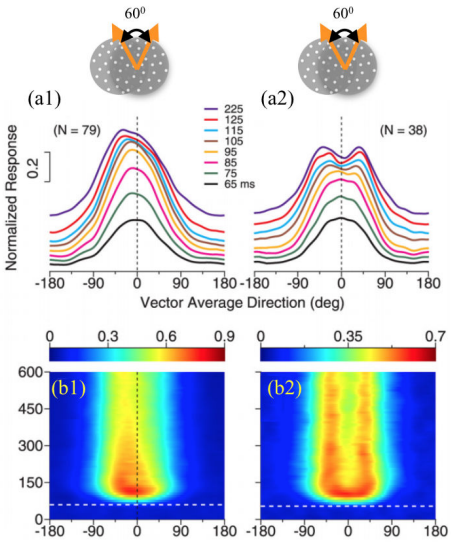


Figure 2.17: Illustrating the temporal evolution of the response selectivities of MT neurons when stimulated with overlapping RDKS, initally the response was observed to be biased towards a Vector Average direction and later on different types of selectivitieis such as side bias of two peaked responses were emergent. Figure adapted from [Xiao 2015].

MT dynamics seems to follow the dynamics of the reported percepts and observed behavioural signatures in terms of eye movements. However, how the population response represents the percept remains ambiguous as the tuning behaviour of the sub-populations such pattern cells, component cells or side-biased cells is different and their role in the eventual percept is not very clear and more over the population responses are obtained by spatial aggregation.

2.2.2.4 Neural representation of coherent and transparent percepts

Eventhough neural correlates of coherent percepts have been found in the form of pattern direction selective cells in MT, it has remained unclear, how the visual system represents transparency? The early conjecture being that pattern direction selective cells represent coherent motion, component direction selective cells could be signalling transparent motion and spatially averaged population responses were considered to reflect the percept. That is, if the spatially averaged population tuning of the MT direction selective cells has two peaked response, it could be signalling transparent motion or if it has a single peak, it could be signalling pattern motion. Early experimental evidence supporting this conjecture came in the form of the experiment by [Snowden 1991], in which enhanced component cell activity

was reported when the cells were stimulated with plaid patterns whose intensity at the intersections were modified to support transparency as illustrated in Fig. 2.18.

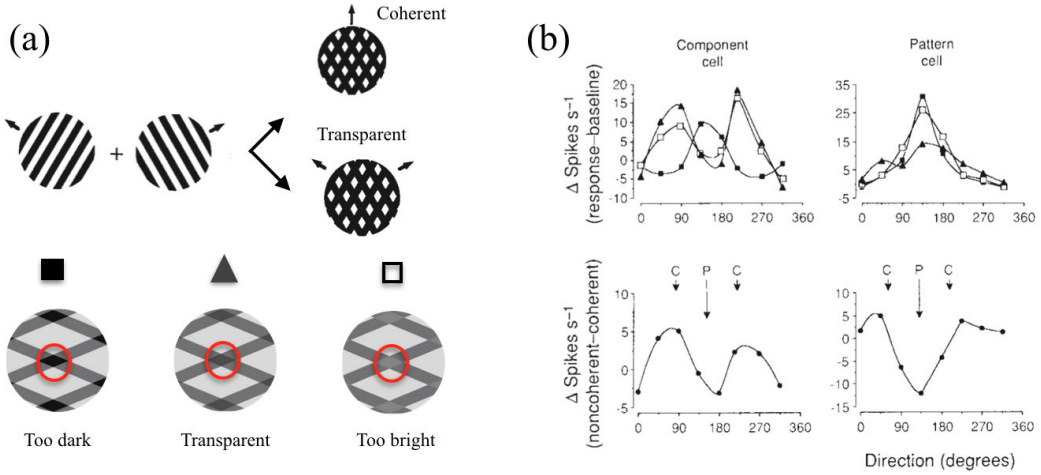


Figure 2.18: Illustrating enhanced activity in component cells when stimulated with plaid patterns evoking a transparent percept. (a) Brightness at the intersections found to play a critical role in percept, if modulated to support the levels that could be expected from physically transparent surfaces, it has been found to evoke perceptual transparency. (b) Responses of pattern and component cells under different conditions of the brightnes at intersections. Figure adapted from [Snowden 1991].

This conjecture was challenged later on by other experiments. [Treue 2000] has found that population response to RDKs giving raise to a transparent percept failed to exhibit bi-modal tuning. [Treue 2000] further suggested that width of the tuning is reflective of the transparency rather than the number of peaks and provided metameric stimuli made up of several different combinations of RDKs as supporting evidence. Different motion stimuli combined such that their overall tuning width is identical were found to be evoking similar percepts as illustrated in Fig.2.19.

This hypothesis was found only relevant for RDKs as [McDonald 2014] observed that, the population response profile for transparent dot fields was narrower than that was observed for coherent plaids, thus one would be unable to decipher the number of motion components perceived directly based on width of the population tuning curves. This lead to further recommendation made by [McDonald 2014] that one needs to take into account the tuning behaviour exhibited by sub-population of the pattern direction selective cells. This claim has been supported by their experimental finding that pattern direction selective cells exhibit uni-modal tuning for plaid stimuli and bi-modal tuning for superimposed RDKs as illustrated in Fig. 2.20.

Countering the observation that pattern cells could play a vital role in decoding percepts no consistent transition in the tuning behavior was found when cells are stimulated with RDKs versus plaids [Xiao 2015], different cells classified as averaging, side biased or two peaked based on their responses to RDKs were found to

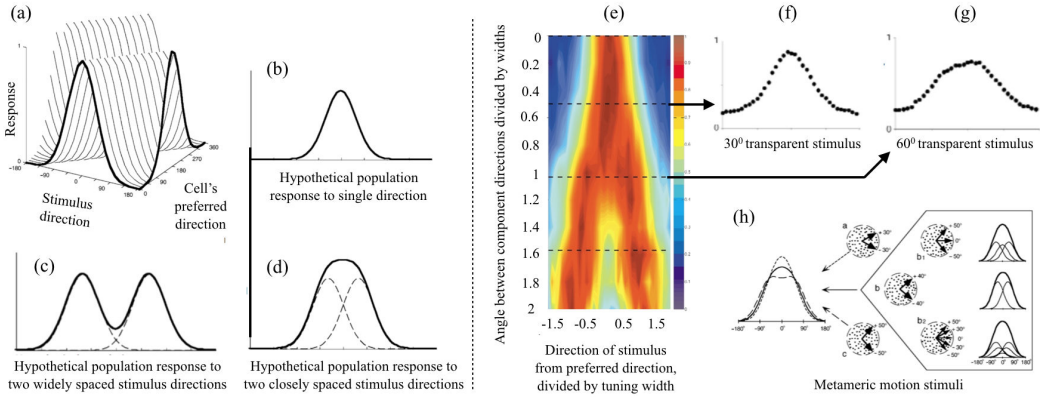


Figure 2.19: Illustrating hypothesis that population tuning width could encode perceptual transparency. (a-d) Revealing hypothetical tuning curves obtained by linear combination of gaussians representative of the component tuning. (e) Population tuning for overlapping RDKs with different degrees of component separation. (f-g) Unimodal population tuning curves observed in case of transparent percepts. (h) Metameric stimuli having identical population tuning obtained via different combinations of components and evoking similar percepts. Figure adapted from [Treue 2000].

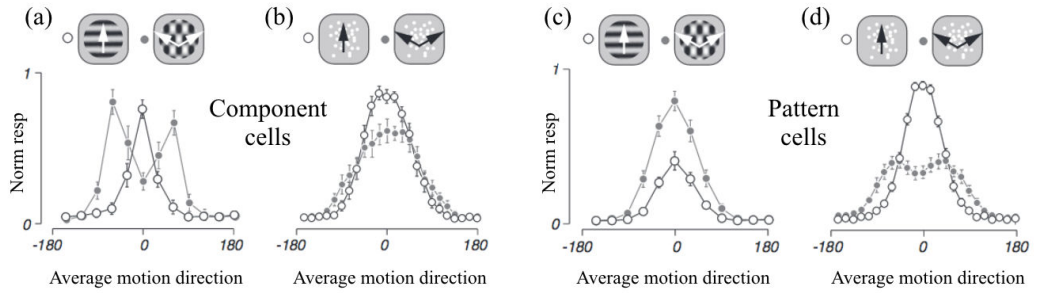


Figure 2.20: Illustrating the unimodal and bimodal tuning exhibited by pattern cells reflective of the percepts evoked with plaid and RDK stimuli. Figure adapted from [McDonald 2014].

exhibit either pattern selectivity on component selectivity when tested with plaid patterns as shown in Fig. 2.21. Also, these sub-groups of direction selective cells have more complex behaviour as tuning curves exhibit a single component selection behaviour in the form of side bias going beyond the uni-model (pattern) or bi-modal (component) selectivity that have been reported so far. These observations raise several questions, what is the contribution of each of these tuning behaviours to an eventual percept? Are changes in tuning behaviour a result of an interplay between driving stimuli and recurrent interactions within the network? Do we need to take a deeper look into the idea of spatially segregating the population responses?

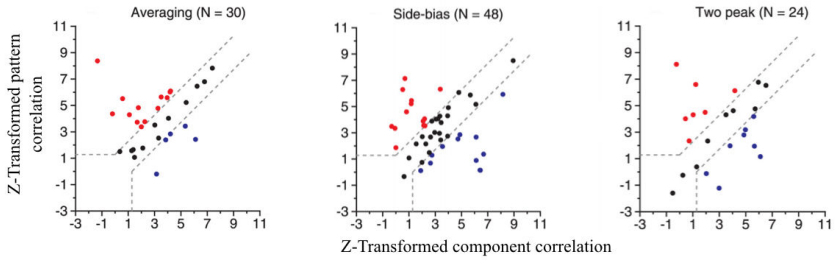


Figure 2.21: Illustrating the behavioural transitions of different sub-populations when stimulated with RDKs versus Plaids. There is no consistent behavioural shift in pattern or component selectivity with plaids to sub-types identified using RDK stimuli. Each graph illustrates the behaviour the neurons classified to be one of the sub-types with RDK stimuli.

2.2.3 Consistent tuning behaviour across stimuli type

One of the characteristics of the pattern motion computation is that velocity estimates are expected be independent of the underlying pattern/image structure. This follows from the observation that once the aperture problem is solved, the cells are expected to be tuned to the velocity features. In physiology, RDKs and plaid patterns are the dominant types of stimuli used to study integration and segregation of motion. So, it is normal to ask whether the velocity tuning behavior exhibited in one class of stimuli translates to the other class thus being able to generalize the hypothesis to complex naturalistic stimuli. However, it is very interesting to notice that very few experiments have been performed testing the tuning behavior of the MT cells with both types of stimuli simultaneously [McDonald 2014, Xiao 2015]. Once, the direction preference of MT cells is established using gratings, experiments are typically performed either with plaid patterns or random dots independenlty, very little is known about the transfer of tuning properties across stimuli types. Do pattern motion selective cells also get tuned to vector average or IoC direction when tested with overlapping RDKs? Recent studies have performed this test with contradicting results [McDonald 2014, Xiao 2015]. Whilst, [McDonald 2014] have shown that response behavior of the pattern cells and component cells is essentially inverted when stimulated with RDKs versus plaids, [Xiao 2015] have found that not only there is no correspondence between the tuning behaviour for pattern and component cells but also it was not possible to predict a consistent transformation from tuning behaviour exhibited for RDKs to one a type of tuning behaviour in plaids. Even though sub-populations exhibiting different kinds of tuning are associated with percepts, lack of consistent tuning across stimulus types needs careful examination as it could signal that observed tunings are result of network effects under different kinds of external stimulations.

To summarize, neural correlates of a solution to the aperture problem have been found in MT area in the form of sub-populations of direction selective cells respond-

ing to pattern motion and the temporal dynamics of the firing seem to qualitatively mimic reported percepts and behavioural responses at different time scales. However, the observations have been deeply coupled with stimuli type used to probe such as RDKs or plaid patterns leaving several important questions open. For us to understand low level vision, which essentially operates in a complex naturalistic environment, we need to arrive at a unified understanding going beyond stimulus categories. In order to do so, we need to characterize the mechanisms that underlie the temporal dynamics and behavioural bifurcations with respect to the stimuli. In the next section, we discuss mathematical models that attempt to capture observations that have been described so far.

2.3 Models of motion estimation

Models of pattern velocity (global velocity) estimation consider it to be a two stage process [Adelson 1982, Grossberg 2001, Bayerl 2004]. In the first stage, local velocities are estimated in the direction orthogonal to local contrast (a minimal contextual attribute) and in the second stage the constituent local velocities are combined to compute the pattern velocity. The models which describe the first stage of motion computation are typically referred to as *motion detection* models and models describing the second stage are know as *motion integration* models. Within the models of motion integration, there are different types, phenomenological models [Adelson 1982, Johnston 2013], which explain how pattern motions are computed based on the constituent components, typically summarizing reported percepts. Second category being those of mechanistic models which try to explain the neural implementation of phenomenological rules [Simoncelli 1998, Bayerl 2004, Tlapale 2011b]. Mechanistic models typically rely on temporal dynamics (behavioral or neuronal) as a signature for cross-validating their efficacy in characterizing the neural mechanisms. In comparison with phenomenological rules, they go beyond the steady state or convergence results often in the form of reported percepts and try to explain how the neural system might potentially arrive at the percept or what processes could be responsible for behavioral dynamics and switches in the percepts. Based on the time scales they consider, these models could be further classified into two types, first type considering early temporal dynamics towards motion integration such as the ones by [Bayerl 2004, Tlapale 2011b] or which focus exclusively on slow scale perceptual switches [Rankin 2013].

2.3.1 Models of motion detection

How does biological vision systems detect motion or how do we build a motion detector? Studies have identified the minimal criteria to be satisfied by a direction selective motion detector [Borst 1989]. The motion detector has to have at least two inputs, a non-linear interaction between the two inputs and asymmetry in the way it process the two inputs. Two inputs are required as a single input cannot successfully

discriminate various scenarios such as temporary change in the illumination or motion in opposite directions. A non-linear interaction is a must to capture temporal attributes of the signal, in the absence of a non-linearity, the time-averaged response of the detector would be equal to the time-averaged input signals thus loosing the information about the temporal sequence of the input. Asymmetry is important to establish direction selectivity, in the absence of the asymmetry one can switch the inputs and the detector response would be same, thus loosing direction selectivity. These three primary components are illustrated in Fig. 2.22.

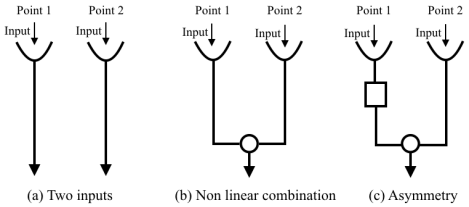


Figure 2.22: Essential components for development of a motion detector, Figure adapted from [Borst 1989].

Several motion detection models were proposed in the literature that satisfy the three criteria listed above [Reichardt 1961, Barlow 1965, D. Marr 1981, Adelson 1985, Van Santen 1985]. Each of them having different roots, for example [Reichardt 1961] was developed to describe insect behavioral studies, [Barlow 1965] was developed to explain retinal physiological observations, [D. Marr 1981] was developed from a computer vision perspective and [Adelson 1985] was based on human psychophysics. These detectors have a lot of common steps and have been broadly classified into correlation based and gradient based [Borst 1989]. Fig. 2.23 illustrates the basic structure of gradient type and correlation type motion detectors. Interestingly correlation based detectors are popular in biological vision whilst gradient based techniques are relied upon in computer vision. A formal equivalence between these different techniques has been already established [Simoncelli 1991]. In this section, we describe an example of each type and then discuss how aperture problem is manifested in these models.

Reichardt Detector: The basic Reichardt detector [Reichardt 1961] comprises of two mirror symmetric units, each of which compute the correlation of time delayed signal at two separate locations. Theoretically, one of those units is sufficient for detection of motion. The correlations in the opposite directions are considered and subtracted from each other (motion opponency) to minimize noise induced correlations. Considering the intensity samples collected by two sensory units located at two spatial locations (x_1, x_2) be denoted by I_{x_1} and I_{x_2}, then the response of a Reichardt detector (R) along the direction $x_1 \rightarrow x_2$ can be given by:

$$R = I_{x_1}(t-\varepsilon) \cdot I_{x_2}(t) - I_{x_1}(t) \cdot I_{x_2}(t-\varepsilon) \tag{2.1}$$

Spatio-temporal energy model or Elaborated Reichardt detectors: The basic structure of the Reichardt detector has been modified later on to include a filtering stage

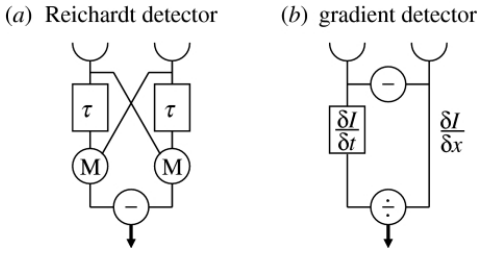


Figure 2.23: Structure of correlation type and gradient type motion detectors. Figure adapted from [Borst 2007].

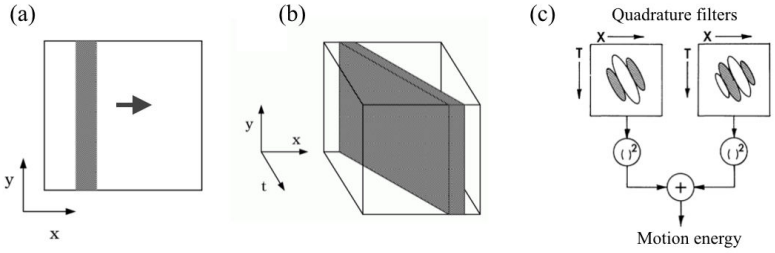


Figure 2.24: Illustrating the idea of motion as a space-time orientation signal. (a) Bar moving horizontally, (b) Moving bar observed in spatio-temporal domain, in resembles an oriented plane in spatio-temporal domain (x-y-t) and an oriented line in x-t domain. (c) Quadrature filters to detect contrast orientation in x-t line, effectively signalling motion.

instead of intensity based correlations to account for contrast dependance on perception [Adelson 1985, Van Santen 1985]. In these extended versions, the output of the basic motion detection unit could be expressed in terms of *"motion energy"*. It is the sum of squared responses of the correlation of input signal with two appropriately chosen quadrature linear filters. The intuition behind these detectors lies in interpreting motion detection problem as space-time contrast orientation detection as illustrated in Fig. 2.24.

Typically, Gabor filters are considered for measuring motion energy. Let F^{odd} and F^{even} be two quadrature Gabor filters given by,

$$F^{odd}(x,t) = exp\left(-\frac{x^2}{2\sigma^2}\right) sin(\omega_x x + \omega_t t) \tag{2.2}$$

$$F^{even}(x,t) = exp\left(-\frac{x^2}{2\sigma^2}\right) cos(\omega_x x + \omega_t t) \tag{2.3}$$

The response of these filters to the spatio-temporal input signal $I(x,t)$ is obtained

$$R_{odd} = I(x,t) * F^{odd}(x,t) \tag{2.4}$$

$$R_{even} = I(x,t) * F^{even}(x,t) \tag{2.5}$$

The local motion energy(M) is extracted by squaring and summing two units outputs. Since, output of two quadrature filters are squared and summed, motion energy is independent of phase.

$$M = \left(R_{odd}^2 + R_{even}^2\right) \tag{2.6}$$

Gradient based Models Instead of the correlation type detectors discussed above, gradient based models operate by relating spatial and temporal changes in the brightness signal [Fennema 1979, D. Marr 1981]. The intuition behind these detectors is that brightness change that is to be expected due to an object motion is proportional to the brightness gradient of the object [Borst 1989]. Considering an intensity pattern $I(x,t)$, this could be expressed as:

$$\frac{\partial I(x,t)}{\partial t} = \frac{\partial I(x,t)}{\partial x} \cdot \frac{dx}{dt} \tag{2.7}$$

The motion velocity along the gradient direction could be expressed as

$$v = -\frac{dx}{dt} = -\left(\frac{\partial I(x,t)}{\partial t}\right) / \left(\frac{\partial I}{\partial x}\right) \tag{2.8}$$

2.3.2 Manifestation of aperture problem in motion detectors

In case of Reichardt detector: The basic Reichardt is supposed to detect motion along the direction in which basic sensory units are aligned. It relies purely based on delayed correlation of the intensity measures, thus it is susceptible misinterpret object motion direction depending on the shape of the object. For example, considering the pair of sensors x_1 and x_2 that are a part of a detector measuring motion in upwards direction as shown in Fig.2.25, depending on the shape of the object, the sensors could be activated even for other motion directions. The detector can only sense the delay between the activation of the two sensors implying that true direction or motion of the object can not recovered.

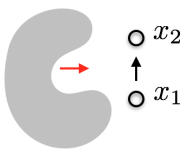


Figure 2.25: Illustrating aperture problem in case of basic Reichardt detector. The sensor could be activated by a variety of motion directions based on the shape of the object.

In case of motion energy models: The motion energy measured by the spatio-temporal oriented filters is a measure of how much the spectrum of the input signal

overlaps or fall with in the spectrum of the filters. Gabor functions that are used to describe these filters have a small, spherical spectrum, whereas the spectrum of a moving object lies along a plane, the orientation of which specifies the object velocity. The filter responds best to planes that pass through the centre of its spectrum, but as planes with various orientation can do this, the filters cannot determine the object velocity as illustrated in Fig. 2.26. This is how aperture manifests itself in the spatio-temporal energy models.

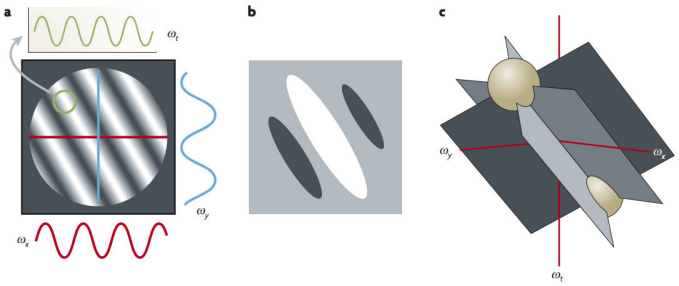


Figure 2.26: Illustrating the manifestation of aperture problem in case of motion energy models. (a) Intuition behind Fourier transform, moving grating stimuli as sinusoidal waves, (b) Example of a filter used to compute motion energy, (c) Spectrum of spatio-temporal filters, spherical blobs indicate filter selectivity, spectrum of the objects with a particular velocity lies along a plane. Figure taken from [Bradley 2008].

In case of gradient based models The velocity information in the gradient models can be derived in closed form by considering the *constant brightness condition*. Let $I(x, y, t)$ denote the measured image intensity at location x,y and time t. Let I_x, I_y, I_t denote the partial derivatives of I with respect to x, y and t, respectively, and let V_x and V_y be the x and y components of the object's velocity. It has been shown that $I_x V_x + I_y V_y = -I_t$, which can be rewritten as $(I_x, I_y).(V_x, V_y) = -I_t$. The intensity derivatives are measurable but we need to estimate V_x and V_y. This equation is ill-posed as we need to recover two variables from one constraint. Thus we can not recover true local velocity of the object but can obtain the estimate of the projection of the velocity in the direction of the gradient.

$$\frac{(I_x, I_y) \,.\, (V_x, V_y)}{\sqrt{(I_x^2 + I_y^2)}} = -\frac{I_t}{\sqrt{(I_x^2 + I_y^2)}} \tag{2.9}$$

2.3.3 Models of motion integration

In the previous section we have seen models of motion detection and manifestations of aperture problem. In this section we visit mechanistic motion integration models which propose solutions to the aperture problem.

2.3.3.1 Mechanistic Models

Simoncelli and Heeger's model: [Simoncelli 1998] proposed a model to explain the pattern direction selectivity of MT neurons following the Fourier domain based interpretation of the V1 cells given by spatio-temporal energy models [Adelson 1982]. The model considers a grey scale video $I(t,x)$: $R^+ \times \Omega \in R^+$ as input, where t denotes time, and $\Omega \in R^2$ is the considered spatial domain. Local contrast is computed $A(t,x) = \left[I(t,x) - \bar{I}(t,x)\right] / \bar{I}(t,x)$, $\bar{I}(t,x)$ is the average of the stimulus over space and time mimicking contrast enhancement by retinal ganglion cells. The model structurally considers feedforward computations by three different sub-population of cells, V1-simple cells, V1-complex cells and MT pattern cells.

- Response of $V1$ simple cells: Let $L(n,t) : R^+ \times 0$ denote the linear filter response to a set of filters along orientations O. $L(n,t) = A \overset{x,t}{*} s_n$, s_n being third derivative of Gaussian filters in 28 different orientations. Then the simple cell responses are given by

$$S(n,t) = \frac{K_1 \lfloor L(n,t)^2 \rfloor}{\sum_m \lfloor L(m,t) \rfloor^2 + \sigma_1} \tag{2.10}$$

  where $\sigma_1, K1$ are empirical constants regulating the saturation response.

- Response of $V1$ complex cells: Obtained by pooling weighted pooling of the responses from various V1 simple cells have same space-time orientation and phase

$$C(n,t) = \sum_m c_{(n,m)} S(m,t) \tag{2.11}$$

- Responses of MT pattern cells: A linear response is computed by selective pooling of the V1 complex cells which lie along the preferred velocity plane followed by a non-linear rectification.

$$q(n,t) = \sum_m p_{nm} C(m,t) \tag{2.12}$$

  where p_{nm} are set of weights that pool the afferent V1 Complex signals. The final response is obtained by rectified divisive normalisation. at the output of the filters, and the output of the MT. It modulates a single cell activity $p(t,i)$ tuned to any given feature i by the population average

$$p(n,t) = \lambda_1 \frac{\lfloor q(t,n) \rfloor^2 |}{\sum_m \lfloor q(t,m) \rfloor^2 + \lambda_2} \tag{2.13}$$

  where $\lfloor x \rfloor = max(0,x)$denotes a positive rectification to account for the firing rate activity, and $\lambda_{1,2}$ are constants.

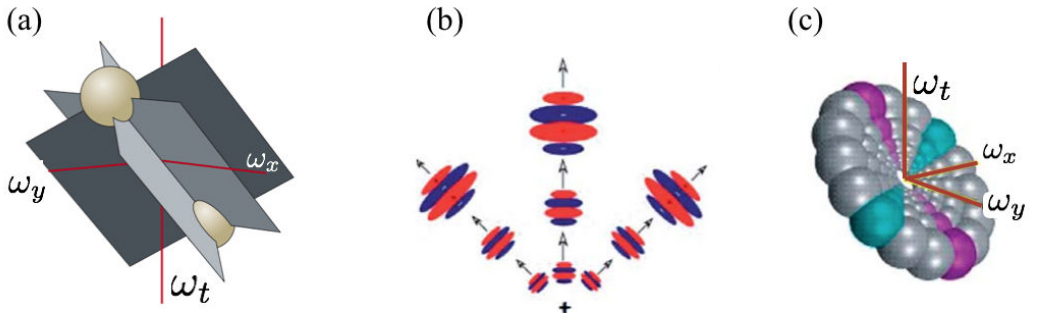


Figure 2.27: Illustrating the pooling strategy to resolve aperture problem. (a) Ambiguity at the level of individual motion energy detectors detectors, (b) Tiling a velocity plane in the frequency domain using an ensemble of filters tuned to different contrast orientations, (c) Pooling the responses along a plane to solve the ambiguity induced by aperture problem. Figure adapted from [Pack 2008, Bradley 2008].

The algorithmic strategy adapted by the model over here is to pool the responses of V1 cells whose response selectivity lies on the plane in the spatio-temporal domain. The ambiguity that is faced by the individual detectors is thus overcome by considering all the evidence that supports a particular object velocity which is done by considering a set of filters that tile the plane as shown in Fig. 2.27(b). This model was shown to be successful in explaining tuning characteristics of MT neurons to gratings and plaid patterns by considering population responses of MT neurons tuned to different velocities sampling the 2D velocity space. However, this model does not take into account the recurrent connectivities between MT neurons and also they do not propose a selectivity scheme as the pooling is spatially isotropic, thus does not deal with scenarios where components have to be selectively integrated such such as occlusions and motion boundaries. Another drawback of this model is that a coherent decoding framework for velocity estimation has not been proposed, making it difficult to test the efficacy of the mechanisms in realistic scenes.

Bayerl and Neumann's model: [Bayerl 2004] proposed that recurrent interactions between V1 and MT can be used to solve the aperture problem by spatial propagation of non-ambiguous cues. Starting from the input image sequence $I : (t, x) \in R^+ \times \Omega \rightarrow I(t, x)$, local motion k_1 is extracted using modified Reichardt detectors. Two filtered images are correlated to estimate population activity: directional derivatives are used to filter the input:

$$c_1(t, x, \alpha) = \frac{I(t, x) \overset{x}{*} \delta_\alpha^2 G_\sigma}{\varepsilon + \sum_{\beta \in O} |I(t, x)| \overset{x}{*} \delta_\beta^2 G_\sigma| \overset{x}{*} G_\sigma} \tag{2.14}$$

where ε avoids divisions by zero, G_σ denotes a Gaussian Kernel, σ's are scaling constants, $\overset{x}{*}$ denotes the convolution operator in space and δ_α^2 denotes the second order directional derivative in the direction $\alpha \in O$

$$c_2^+(t,x,v) = \left(\sum_{\alpha \in O} c_1(t,x,\alpha) c_1(t+1, x+v, \alpha) \right) \overset{x}{*} G_\sigma \tag{2.15}$$

$$c_2^-(t,x,v) = \left(\sum_{\alpha \in O} c_1(t+1,x,\alpha) c_1(t, x+v, \alpha) \right) \overset{x}{*} G_\sigma \tag{2.16}$$

The half detectors are then combined by

$$k_1 t, x, v = \frac{|c_2^+(t,x,v)|_+ - \frac{1}{2}|c_2^-(t,x,v)|_+}{1 + |c_2^-(t,x,v)|_+} \tag{2.17}$$

where $|x|_+ = \max(0,\text{x})$ is a positive rectification, for the activity of neurons is always positive.

The population activity of the two cortical areas, V1 and MT be denoted by p_i,

$$p_i : (t,x,v) \in R^+ \times \Omega \times V \rightarrow p_i(t,x,v) \in [0,1] \tag{2.18}$$

where V represents the space of possible velocities. Each function p_i can be interpreted as the state of cortical area retinotopically organised which describe at each position x the instantaneous activity of a neuron tuned for the velocity V.

$$m_1(t,x,v) = k_1(t,x,v)(1 + 100 p_2(t,x,v)) \tag{2.19}$$

$$n_1(t,x,v) = m_i^2(t,x,v) \overset{*}{v} G_\sigma \tag{2.20}$$

$$p_1(t,x,v) = \frac{n_1(t,x,v) - \frac{1}{2|V|} \sum_{w \in V} n_1(t,x,w)}{0.01 + \sum_{w \in V} n_1(t,x,w)} \tag{2.21}$$

$$n_2(t,x,v) = p_1^2(t,x,v) \overset{*}{x,v} G_\sigma \tag{2.22}$$

$$p_2(t,x,v) = \frac{n_2(t,x,v) - \frac{1}{2|V|} \sum_{w \in V} n_2(t,x,w)}{0.01 + \sum_{w \in V} n_2(t,x,w)} \tag{2.23}$$

where m_i and n_i are intermediate stages to compute pi_i, k_1 is the local motion input. $\overset{*}{v}$ denotes convolution with respect to the spatial domain and with a Gaussian whose standard deviation is o.75 and $\overset{*}{x,v}$ denotes convolution with respect to the spatial and velocity domains with a Gaussian whose standard deviation is 0.75 (velocity) and 7 (spatial).

The algorithmic strategy adapted by this model is to solve the aperture problem by propagation of unambiguous cues from spatially distant regions using strong feedback from MT cells with large receptive fields. This is achieved by considering

a form independent local velocity distribution that encodes uncertainty at the V1 stage. The regions with 2D cues generate responses with strong localized peaks and ambiguous regions generate broad low amplitude distributions. The strong shunting inhibition sets the network in winner take all mode, allowing enhanced activity for the non-ambiguous cues arriving via modulatory feedback from MT, thus solving the aperture problem. This is illustrated in Fig. 2.28. This idea has been very influential and has been widely considered by other models [Beck 2011, Tlapale 2011b, Raudies 2011].

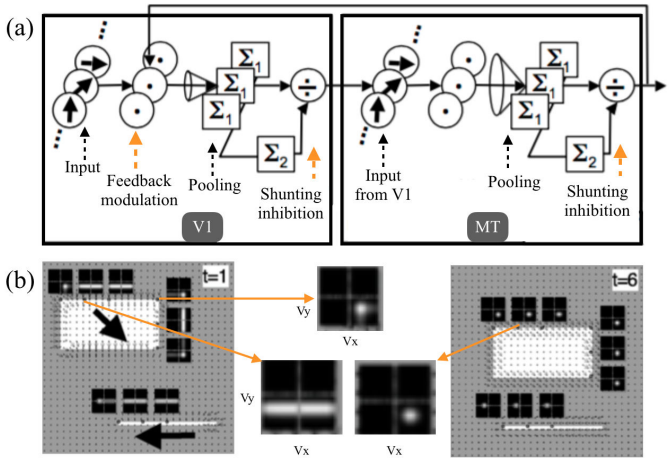


Figure 2.28: Illustrating the model of motion integration proposed by [Bayerl 2004]. (a) Computational steps describing processing at V1 and MT stages, (b) Showing unambiguous motion detected at corners and resolution of the ambiguous motion estimation along the edge by the model. Figure adapted from [Bayerl 2004].

However, this model ignores few important aspects. At convergence, the activity of V1 and MT are identical, thanks to the strong feedback coupling. V1 cells have not been reported to exhibit such a dynamic change in their tuning behaviour. The second aspect is that a strong feature domain inhibition is considered thus eliminating the possibility of transparency and only allows modelling of pattern selective cells. The third aspect is also that, the model relies on explicit 2D cues available at the V1 stage, thus focussing on the spatial propagation aspects of the aperture problem. The model does not give much attention to the dynamics of both motion integration as well as the interplay between the tuning properties of the selected units and recurrent interactions.

Florian Raudies and Ennio Mingolla's model: [Raudies 2011] proposed an extension to the model by [Bayerl 2004] to capture motion transparency where multiple motions are perceived at a single spatial location. This model proposes center-surround interactions in the velocity domain would allow for co-existence of multiple activity peaks in MT population responses and thus can capture transparency. The model is quite complex and considers recurrent interactions between several different

cortical areas: V1, MT, MST, LIP, each of the stages sharing common computational steps. Here, we would only discuss some key ideas and limitations instead of giving a detailed mathematical description of these stages. The model rightly points out that center-surround interactions are critical in capturing transparency going beyond the strong shunting inhibition that is typically considered. The center-surround connectivities are chosen such that the population tuning of MT cells closely resembles the experimental reportings by [Treue 2000]. Co-existing activity peaks in the velocity domain are obtained by truncating the overall interaction widths. The model also raises another important issue that in the literature spatially dense decoding has not been attempted to explain transparency, instead population responses are pooled across the images and analysed. However, the model does not present results of spatially dense decoding. Due to the nature of the kernels considered the model focusses on explain transparency in large stimuli such as patches of random dot stripes and does not go into details in cases of stimuli such as plaids.

2.4 Discussion and Conclusion

Both physiological and psychophysical experiments explore simplistic stimuli in order to explore the mechanisms of motion integration. While psychophysical experiments have elucidated phenomenological rules of cue combination, physiological experiments have identified possible neural correlates. However, these rules are formulated on the basis of simplified stimuli in specific conditions and often the models describing the results are only tested with limited sub set of stimuli. The focus of the models has not been on recovering the dense optical flow field, so, it is not clear how the models would be performing in complex realistic scenes and what kind of improvements need to be taken into account do deal with the complexities.

The first step in developing a scaled up models would be to identify the cell types to be included. Psychophysical observations of the coherent and transparent percepts have been weakly linked to different cell types, the pattern cells and component cells which exhibit uni-modal and bi-model tuning respectively. But considering this kind of static tuning behaviour by cells has been questioned by recent experimental observations by [Xiao 2015], who showed that tuning behaviours don't translate across stimuli. This raises a serious problem of what are the minimal cell types to be included in a model to explain percept or estimation of optical flow by integration and segregation? Also, including more and more cell types with distinct tuning behaviour to account for experimental observations might dangerously lead to data over fitting and resulting in less insights into the underlying network computation.

Psychophysical observations at multiple scales using stimuli such as moving gratings and lines under apertures hint at spatial propagation and competition between different motion cues and which is also reflected in the form of dynamic shifts in the neuronal selectivity. In feed forward models, such as [Perrone 2008], this kind of shifts are accounted for by adding additional inputs to the MT cells, either tem-

porally delayed or in the form of cues from form related areas. But, it fails to explain the lack of consistent transition in the tuning behaviour when the driving input is changed. Both, the dynamic shifts in the tuning behaviour and lack of consistent transition from one stimulus type to other hint a strong role of recurrent interactions. Even though there are spatialized recurrent network models [Beck 2011, Tlapale 2011b, Raudies 2011] that have been proposed in the literature, these models consider 2D cues to be either computed in the afferent stages or rely on empirically chosen connectivities and do not analyse the interplay between tuning behaviour and driving stimuli. Thus they break a continuum between local 1D cue integration and form based spatial propagation of the activity. Thus we need to better understand the local recurrent interactions that could facilitate spatio-temporal grouping by direction selective MT units. In thesis, I examine the following questions:

- What is the efficacy of the models proposed in psychophysics/physiology in dealing with real scenes?
- How to improve the performance of the models?
- How does one explain dynamic shifts in tuning properties?
- How does one capture transitions in the tuning behaviour across stimuli type?

Part II

Feedforward models for motion estimation

Chapter 3

What can we expect from a feedforward V1-MT model?

"Testing leads to failure and failure leads to understanding."

- Burt Rutan, *Aviation pioneer*

In this chapter, our goal is to scale up a biologically inspired motion estimation model and benchmark it on a state-of-the-art computer vision dataset. In order to do so, we focus on the V1-MT feedforward model, which is minimal and can be considered equivalent to the popular and well studied Lucas-Kanade approach [Lucas 1981] (see [Simoncelli 1991]).The two key contributions of this work can be stated as follows:

- Proposing a velocity space sampling of tuned MT neurons and a scheme to decode the local velocity from the activity of these neurons. Estimating spatially dense flow field in contrast to recovering a single readout.
- Examining the efficacy of V1-MT feedforward processing in natural image scenarios. The stimuli used in various psychophysical and physiological experiments that inspired the V1-MT feedforward model are highly homogeneous and do not cover the complexities that arise in the case of natural images [Nishimoto 2011]. Thus the efficacy of the proposed model and inherent limitations in case of natural stimuli are not known. This is explored by considering Middlebury dataset, which comprises complex stimuli.

In the past two decades, efforts by computer vision researchers have led to development of a large number of models for the computation of optical flow (see [Fortun 2015] for a review). In addition to modelling efforts to solve this task, a prominent achievement in computer vision has been to develop publicly available benchmarking datasets [Baker 2011, Butler 2012] to evaluate and compare models in natural image scenarios. These benchmarking datasets have spurred a great deal of research resulting in new models, however, despite this large amount of work in this area, the problem still remains open as many of the models either lack consistent accuracy across video sequences or have a high computational cost.

On the other hand the neural mechanisms underlying motion analysis in the visual cortex have been extensively studied with a lot of emphasis on understanding the function of cortical areas V1 [Sincich 2005, Rasch 2013] and MT [Rust 2006], which play a crucial role in motion estimation (see [Perrone 2008, Bradley 2008,

Pack 2008] for reviews). Neurons in V1 are found to respond when motion direction is perpendicular to the contrast of the underlying pattern, while neurons in MT are found to respond best to a particular speed irrespective of the underlying contrast orientation and thus are believed to be solving the local motion estimation problem.

Several computational models have been proposed based on the available experimental data. Initially models focussed on motion sensitive cells in V1 (complex cells). Using the conceptual framework of receptive fields (RF) the responses were explained using Gabor functions [Daugman 1985], and spatio-temporal motion energy [Adelson 1985]. Then few attempts were made to recover the motion vectors directly from the motion energy representation [Heeger 1987, Grzywacz 1990]. One could call these models as being at the interface between computer vision and biological vision. These initial attempts were later on leveraged and extended to explain the properties of MT neurons by considering a feedforward pooling from V1 cells followed by divisive normalisation [DeAngelis 1995, Simoncelli 1998, Rust 2006]. Apart from this class of linear-non linear feedforward models other attempts were made to simulate the information processing by V1-MT layers using lateral or feedback interactions for solving the aperture problem, by considering a pure velocity space representation and various kinds of local motion estimation [Bayerl 2004, Bayerl 2007b, Tlapale 2010, Masson 2010, Bouecke 2011].

Even though there was some early interaction among the biological and computer vision communities at a modeling level (see, e.g., [Heeger 1988, Nowlan 1994, Simoncelli 1998]), comparatively little work has been done for examining or extending the models proposed in biology in terms of their engineering efficacy on modern optical flow estimation datasets. In this work, we take a step towards filling the critical gap between biological and computer vision studies (see [Medathati 2015b] for a more general discussion), focusing on visual motion estimation, leveraging and testing ideas proposed in biology in terms of building scalable algorithms.

The chapter is organized as follows: In Sec. 3.1 we present our V1-MT feedforward architecture for optical flow estimation (called FFV1MT). Our model has three main steps: The two first steps model V1 cells and MT pattern cells following classical ideas from the literature. The third step is a decoding stage to extract the optical flow from MT population response. In Sec. 3.2 we present the algorithmic details of this model, which are an essential contribution here, since they allow this V1-MT architecture to be applied to real videos. In particular, we propose a multi-scale approach to deal with large ranges of speeds found in natural scenes. In Sec. 3.3 we evaluate our approach on several kinds of videos. We use test sequences to show the intrinsic properties of our approach and we benchmark our approach using the Middlebury dataset [Baker 2011].

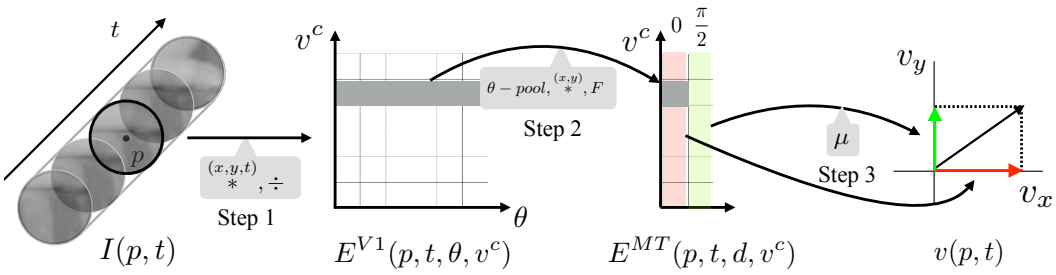


Figure 3.1: FFV1MT Model overview: It is a three-step feedforward model, where Step 1 corresponds to the V1 layer (obtained by a non-separable spatio-temporal filtering and a normalisation), Step 2 corresponds to MT layer (obtained by pooling V1 responses first with respect to θ, then in a local spatial neighbourhood, and applying a static non-linearity) and Step 3 is velocity estimation (obtained by a weighted average of MT responses).

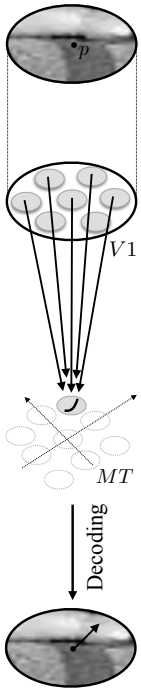
3.1 Feedforward V1-MT model for optical flow estimation

3.1.1 General overview

In general, the pattern selectivity of MT cells can be explained by following two different approaches [Bradley 2008]: the motion computation can be related to direct 2-D feature tracking mechanisms, or based on fusion of 1-D velocity cues using intersection of constraints (IOC) mechanisms. For the former approach, the consequence is that the aperture problem does not affect the motion processing, though little evidence for a feature-tracking mechanisms are reported [Stoner 1990, Noest 1993, Skottun 1999]. The latter approach is based on geometric relationships among the local velocity estimates.

The model we study in this chapter is based on a non-linear integration of the V1 afferents to obtain the MT pattern cells [Pack 2008]. In particular, the IOC mechanism is indirectly considered through localized activations of V1 cells [DeAngelis 1995, Simoncelli 1998, Rust 2006]. It is a three-step feedforward model: Step 1 corresponds to the V1 simple and complex cells, Step 2 corresponds to the MT pattern cells and Step 3 corresponds to a decoding stage to obtain the optical flow from the MT population response. In term of modelling, Steps 1 and 2 follow a classical view, while Step 3 has been introduced to solve the task of optical flow. An illustration of our model called FFV1MT is given in the figure next to Tab. 3.1 (see also Fig. 3.1 for a more detailed illustration of the computations involved).

This model is inspired from previous works from visual neuroscience [Heeger 1987, Simoncelli 1998, Rust 2006] and in Tab. 3.1, we summarise the main differences. In the seminal paper of Heeger [Heeger 1987] a first motion estimation model is introduced to compute the optical flow. Steps 1 and 2 of



Model characteristics	[Heeger 1987]	[Simoncelli 1998]	[Rust 2006]	FFV1MT
V1 cell model	Gabor filters	Third derivative of a Gaussian	Direction space only	Gabor filters as in [Heeger 1987]
MT pooling	N.A.	Yes	Yes	Yes
MT nonlinearity	N.A.	Yes	Yes	Yes
MT population sampling	N.A.	Dense	Direction space only	Principal axes only
Decoding	Least-square on motion energy	No	No	Linear
Multi scale	Yes	No	No	Yes
Coarse-to-fine	No	No	No	Yes

Table 3.1: Comparison of our model FFV1MT with respect to other most related work.

our model are similar to the ones presented in [Simoncelli 1998], but in the latter the optical flow is not estimated. It is worth to note that the model proposed in [Rust 2006] is described in the parameter space, whereas we present a model in the (p, t) space that is able to estimate the optical flow of real-world sequences. All the models, but [Rust 2006], introduce a processing stage to avoid responses to ambiguous low frequency textures. Finally, we propose an empirical sampling scheme of the two-dimensional velocity space, which provides competitive estimates while reducing the computational cost significantly when compared to [Simoncelli 1998].

3.1.2 Description of the FFV1MT model

Let us consider a grayscale image sequence $I(p, t)$, for all positions $p = (x, y)$ inside a domain Ω and for all time $t > 0$. Our goal is to find the optical flow $v(p, t) = (v_x, v_y)(p, t)$ defined as the apparent motion at each position p and time t.

Step 1 : V1 (Motion energy estimation and normalization) In the V1-layer two sub-populations of neurons are involved in the information processing, namely V1-direction selective simple cells and complex cells. Simple cells are characterised by the preferred direction θ of their contrast sensitivity in the spatial domain and their preferred velocity v^c in the direction orthogonal to their contrast orientation often referred to as component speed. The RFs of the V1 simple cells are classically modelled using band-pass filters in the spatio-temporal domain. In order to

achieve low computational complexity, the spatio-temporal filters are decomposed into separable filters in space and time. Spatial component of the filter is described by Gabor filters $\mathcal{H}$ and temporal component by an exponential decay function $\mathcal{P}$. Given the peak spatial and temporal frequencies f_s and f_t of a receptive field, we define the following complex filters by:

$$\mathcal{H}(p,\theta,f_s) = Be^{\left(\frac{-(x^2+y^2)}{2\sigma^2}\right)}e^{j2\pi(f_s cos(\theta)x + f_s sin(\theta)y)}, \tag{3.1}$$

$$\mathcal{P}(t,f_t) = e^{\left(-\frac{t}{\tau}\right)}e^{j2\pi(f_t t)}, \tag{3.2}$$

where σ and τ define the spatial and temporal scales, respectively. Denoting the real and imaginary components of the complex filters $\mathcal{H}$ and $\mathcal{P}$ as $\mathcal{H}_e, \mathcal{P}_e$ and $\mathcal{H}_o, \mathcal{P}_o$ respectively, and a preferred velocity v_c related to the frequencies by the relation

$$v^c = \frac{f_t}{f_s}, \tag{3.3}$$

we introduce the odd and even spatio-temporal filters defined as follows,

$$\begin{aligned}\mathcal{G}_o(p,t,\theta,v^c) &= \mathcal{H}_o(p,\theta,f_s)\mathcal{P}_e(t,f_t) + \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_o(t,f_t),\\ \mathcal{G}_e(p,t,\theta,v^c) &= \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_e(t,f_t) - \mathcal{H}_o(p,\theta,f_s)\mathcal{P}_o(t,f_t).\end{aligned} \tag{3.4}$$

These odd and even symmetric and tilted (in space-time domain) filters characterize V1 simple cells. Using these expressions, we define the response of simple cells, either odd or even, with a preferred direction of contrast sensitivity θ in the spatial domain, with a preferred velocity v^c and with a spatial scale σ by

$$R_{o/e}(p,t,\theta,v^c) = (\mathcal{G}_{o/e}(\cdot,\cdot,\theta,v^c) \overset{(x,y,t)}{*} I)(p,t). \tag{3.5}$$

Fig. 3.2(a) shows the amplitude power spectra of the spatio-temporal filters $\mathcal{G}_o(p,t,\theta,v^c)$ (the same is for $\mathcal{G}_e(p,t,\theta,v^c)$) in the frequency domain. The shape of the amplitude power spectra of the filters' bank is due to the combination of the odd and even functions ($\mathcal{H}_o$, $\mathcal{H}_e$, $\mathcal{P}_o$, and $\mathcal{P}_e$) given in (3.4).

The complex cells are described as a combination of the quadrature pair of simple cells (5.1) by using the motion energy formulation

$$E(p,t,\theta,v^c) = R_o(p,t,\theta,v^c)^2 + R_e(p,t,\theta,v^c)^2,$$

followed by a normalisation: Considering a finite set of orientations $\theta = \theta_1 \ldots \theta_N$, the final V1 response is defined by

$$E^{V1}(p,t,\theta,v^c) = \frac{E(p,t,\theta,v^c)}{\sum_{i=1}^{N} E(p,t,\theta_i,v^c) + \varepsilon}, \tag{3.6}$$

where $0 < \varepsilon \ll 1$ is a small constant to avoid divisions by zero in regions with no energy (when no spatio-temporal texture is present). The main property of V1 is its tuning to the spatial orientation of the visual stimulus, since the preferred velocity of each cell is related to the direction orthogonal to its spatial orientation.

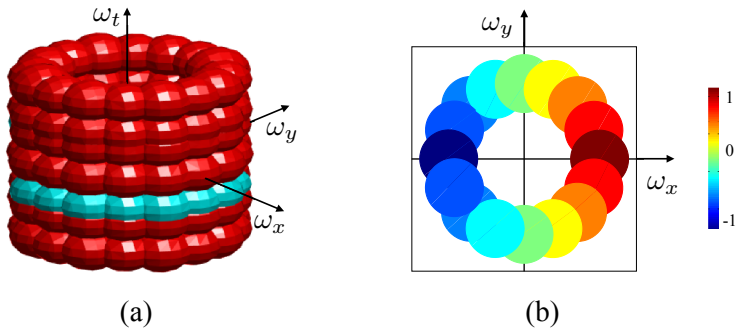


Figure 3.2: Representation of the V1 RFs in the frequency domain. (a) The iso-surface of the power spectra of the considered spatio-temporal filter bank that models the V1 cells. The spatial radial peak frequency of the filters is constant and the temporal frequency changes, thus the frequency bands have a cylinder-like shape. The V1 cells afferent to a population of MT cells for a specific v^c are highlighted in cyan. (b) The weights $w_d(\theta)$ used to pool the afferent V1 cells. In particular, the weights refer to a cosine weighting function, with values from -1 to 1 as in the colormap.

Step 2: MT pattern cells response MT neurones exhibit velocity tuning irrespective of the contrast orientation. This is believed to be achieved by pooling afferent responses in both spatial and orientation domains followed by a nonlinearity [Simoncelli 1998]. The responses of an MT pattern cell tuned to the speed v^c and to direction of speed d can be expressed as follows:

$$E^{MT}(p,t,d,v^c) = F\left(\sum_{i=1}^{N} w_d(\theta_i) G_{\sigma_{pool}} \overset{x,y}{*} E^{V1}(p,t,\theta_i,v^c)\right), \tag{3.7}$$

where $G_{\sigma_{pool}}$ denotes a Gaussian kernel of standard deviation σ_{pool} for the spatial pooling, $F(s) = exp(s)$ is a static nonlinearity chosen as an exponential function [Paninski 2004, Rust 2006], and w_d represents the MT linear weights that give origin to the MT tuning. In Fig. 3.2(a) the power spectra of the filters corresponding to the V1 cells afferent to a population of MT cells tuned to a specific v^c are represented in cyan. Such afferent cells are weighted through the $w_d(\theta)$, as shown in Fig. 3.2(b).

Physiological evidence suggests that w_d is a smooth function with central excitation and lateral inhibition. Cosine function shifted over various orientations is a potential function that could satisfy this requirement to produce the responses for a population of MT neurones [Maunsell 1983a]. Considering the MT linear weights shown in [Rust 2006], $w_d(\theta)$ is defined by

$$w_d(\theta) = \cos(d-\theta) \quad d \in [0, 2\pi[. \tag{3.8}$$

This choice allows to obtain direction tuning curves of pattern cells that behave as in [Rust 2006]. However, considering MT neurones that span over the 2-D velocity

space with a preferred set of tuning speed directions in $[0, 2\pi[$ and also a multiplicity of tuning speeds is not necessary to encode velocity. A sampling along the cardinal axes is sufficient to recover the full velocity vector: since cosine functions shifted over various orientations (see Eq. (3.8)) can be described by the linear combination of an orthonormal basis (i.e., sine and cosine functions), all the V1 afferent information is encoded by two populations of MT neurons (see Eq. (5.1.2)). For this reason, in this model, we sample the velocity space using two MT populations tuned to the directions $d = 0$ and $d = \pi/2$ with varying tuning speeds.

Step 3: Decoding In this step we wonder how optical flow can be estimated by decoding the population responses of the MT neurones. Indeed, a unique velocity vector cannot be recovered by activity of a single velocity tuned MT neurone as multiple scenarios could evoke the same activity, but unique vector can be recovered based on the activity of a population. In this chapter, we present a decoding step which was not present in [Simoncelli 1998, Rust 2006] to decode the MT population. We adopt a linear combination approach to decode the MT population response as in [Pouget 1998, Rad 2011]:

$$\begin{cases} v_x(p,t) = \sum_{i=1}^{M} v_i^c E^{MT}(p,t,0,v_i^c), \\ v_y(p,t) = \sum_{i=1}^{M} v_i^c E^{MT}(p,t,\pi/2,v_i^c). \end{cases} \tag{3.9}$$

3.1.3 An extension to deal with discontinuities: The FFV1MT–TF model

The FFV1MT approach described in this section relies on isotropic spatial smoothing at V1 level and isotropic pooling from V1 to MT. There is no mechanism to deal with motion discontinuities. In this section, we propose a simple extension of the FFV1MT model to show how discontinuities could be preserved. The idea is to introduce an iterative diffusion process between MT cells, which could be interpreted as the effect of lateral connections inside the MT population. The way nearby cells exchange information depends on their respective tuning speeds and directions, but it can also depend on the local context of the image. For example, local contrast and luminance information can modulate neurones characteristics and connections.

To model this idea, we propose a solution based on the trilateral filter (TF) which is an extension of the linear Gaussian filtering. Bilateral and trilateral filter have been extensively used in the context of nonlinear image smoothing leading to many applications (see [Paris 2009] for a review). They provide a simple way to take discontinuities into account. Considering each population of MT cells tuned to a specific value of d and v^c as a spatial map, the goal is to apply TF in space to each map $E^{MT}(\cdot, t, d, v^c)$. This model is called FFV1MT–TF.

Denoting $E^{MT}(p, t, d, v^c)$ by $E^{MT}(p)$ for sake of simplicity, one iteration of TF

on $E^{MT}(p)$ is defined by:

$$TF_{\alpha,\beta,\gamma}[E^{MT}](p) = \frac{1}{N(p)} \int_{p' \in \Omega} f_\alpha(\|p - p'\|) f_\beta(E^{MT}(p') - E^{MT}(p)) \\ f_\gamma(I(p',t) - I(p,t)) E^{MT}(p') dp', \tag{3.10}$$

where

$$f_\mu(s) = \exp(s^2/\mu^2) \quad s \in \mathbb{R}, \tag{3.11}$$

α, β and γ are parameters defining the smoothing properties of TF and $N(p)$ is the normalising term

$$N(p) = \int_{p' \in \Omega} f_\alpha(\|p - p'\|) f_\beta(E^{MT}(p') - E^{MT}(p)) f_\gamma(I(p',t) - I(p,t)) dp'.$$

The interpretation of (3.10) is that, to estimate the new activity of an MT cell located at position p after one pass of TF, we average MT cell activities which are close in space, which have a similar activity, and which correspond to positions having similar luminance. The resulting filtered energy $TF_{\alpha,\beta,\gamma}[E^{MT}](p)$ is smoothed while main discontinuities are preserved and enhanced according to energy and luminance discontinuities. Several iterations of this filter can be made depending on the degree of smoothing desired.

3.2 Making the approach applicable to real videos

This kind of V1-MT feedforward architecture presented in Sec. 3.1 was initially proposed to explain recorded neural activities and mainly applied on synthetic homogeneous images such as moving gratings and plaids. They were not designed to be a systematic alternative to computer vision algorithms to work on real videos. In this section, we propose algorithmic solutions to make this V1-MT feedforward architecture applicable to real videos so that it could be benchmarked using state-of-the-art dataset.

3.2.1 Multiscale approach

One critical point in dealing with real videos is to be able to deal with a large range of speeds. As detailed in Sec. 3.1, the V1-like RFs are modelled through spatio-temporal filters. In order to keep as low as possible the computational load of the model, only one spatial radial peak frequency f_s has been considered. This is in contrast with the physiological findings, since information in natural images is spread over a wide range of frequencies, it is necessary to use a mechanism that allows to get information from the whole range of frequency.

We propose a multi-scale approach as illustrated in Fig. 3.3. This is a classical approach used in computer vision. It consists in (i) a pyramidal decomposition with L levels [J.R. Bergen 1984] and (ii) a coarse-to-fine refinement [Simoncelli 1993],

which is a computationally efficient way to take into account the presence of different spatial frequency channels in the visual cortex and their interaction.

Using this approach, the spatial distance between corresponding points is reduced, thus yielding to a more precise estimate, since the residual values of the velocities lie in the filters' range. This also allows large displacements to be estimated which is a crucial aspect when dealing with real sequences. Interestingly, at a functional level, there is an experimental evidence that MT neurons seems to follow a coarse-to-fine strategy [Pack 2001] suggesting that motion signals become more refined over time.

The equivalence between a multi-scale approach and the corresponding multi-resolution approach is shown in Fig. 3.4. The multi-scale analysis is performed by using three banks of Gabor filters with different spatial peak radial frequencies, each separated by an octave scale. The multi-resolution approach is obtained by iteratively low-pass filtering and sub-sampling the input image, then only the outermost bank of filter (i.e., the highest frequency one) is applied.

3.2.2 Boundary conditions

The problem of boundary conditions arises as soon as we need to consider values outside the domain of definition Ω. Even with simple Gaussian smoothing, when estimating results close to the boundaries, one needs to access values outside Ω. This is solved generally by choosing some boundary conditions like Neumann or Dirichlet. However, in our case, using such assumptions might introduce some strong errors at the boundaries. For this reason, we proposed instead to work inside an inner region denoted by Ω_{in} in which only available values are taken into account (so that no approximation or assumption has to be made), and then to interpolate values in the remaining outer region denoted by Ω_{out}. Note that this is an important issue to consider, especially because we use a multi-scale approach since errors done at the boundaries at low scales can spread a lot as scales are getting finer.

The way to defined the outer region Ω_{out} is illustrated in Fig. 3.5(a). It is constructed by first taking into account the region $\mathcal{B}_1$ in which V1 cells would need values outside Ω, and then the regions $\mathcal{B}_2$ corresponding to MT cells that would pool information from V1 cells in $\mathcal{B}_1$. So we have $\Omega_{out} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $\Omega_{in} = \Omega \backslash \Omega_{out}$. Given this definition of inner and outer regions (Fig. 3.5(b)), the idea is to make all the estimations in Ω_{in} and to interpolate values in the outer region Ω_{out} (Fig. 3.5(c)). Given E^{MT} estimated in Ω_{in}, we propose that

$$E^{MT}(p) = \frac{1}{N(p)} \int_{p' \in \mathcal{A}} f_\alpha(\|p - p'\|) f_\gamma(I(p) - I(p')) E^{MT}(p') dp' \quad \forall p \in \Omega_{out}, \quad (3.12)$$

where $\mathcal{A}$ contains pixels at the inner boundary of Ω_{in} (green region) where E^{MT} is well estimated, function f_μ is defined as in (3.11), α and γ are parameters and $N(p)$ is a normalizing term

$$N(p) = \int_{p' \in \mathcal{A}} f_\alpha(\|p - p'\|) f_\gamma(I(p) - I(p')) dp'.$$

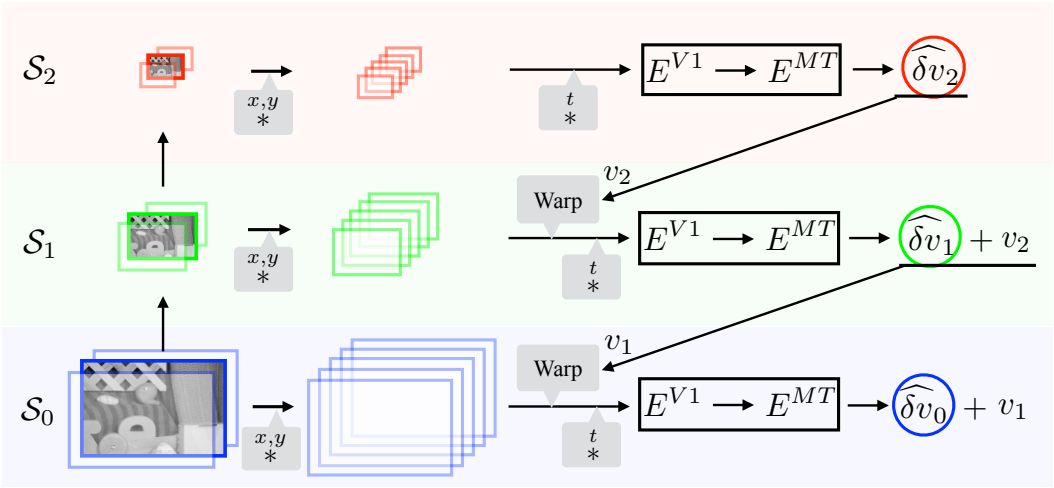


Figure 3.3: Multi-scale approach: In this example, three scales are represented ($L = 3$). Pyramidal decomposition is denoted by $\mathcal{S}_l$ with ($l = 0 \ldots L-1$) ($l = 0$ is the finer scale). At a scale l, the estimated residual optical flow ($\widehat{\delta v_l}$) plus the optical flow coming from the coarser scale (v_{l+1}) is used to warp the sequence of the spatially filtered images at scale $l-1$.

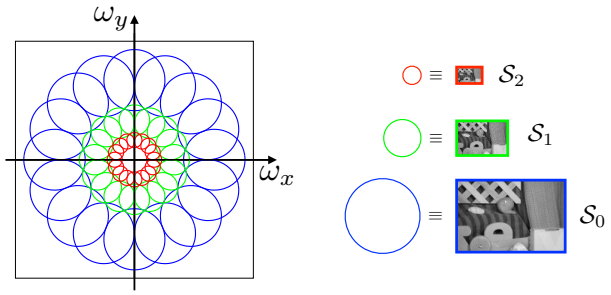
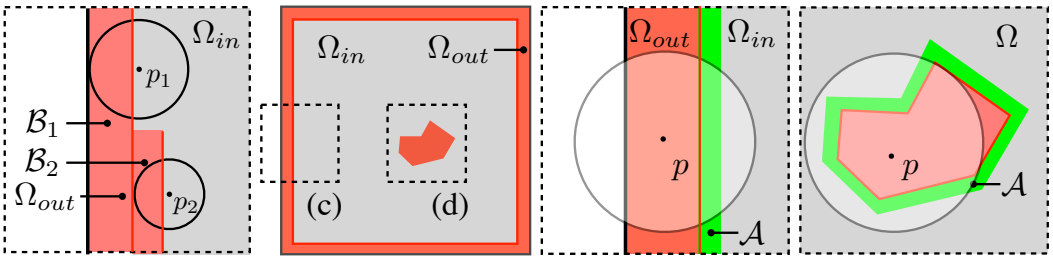


Figure 3.4: Equivalence between a multi-scale approach and the corresponding multi-resolution approach. This figure shows the amplitude spectra of three banks of Gabor filters with three spatial peak radial frequencies and eight spatial orientation: this frequency representation is a slice obtained for a fixed ω_t, the $(\omega_x, \omega_y, \omega_t)$ amplitude spectra of the bank of filters is shown in Fig. 3.2. Processing the image at full resolution by using the three banks of filters is equivalent to apply the outermost bank of filters to the three sub-sampled images.

This method is based on luminance similarities using the same idea as developed in Sec. 3.1.3. Note that other interpolation methods could be used instead.

3.2.3 Unreliable regions

A problem is found with regions having a null spatio-temporal content, which happens for example in the blank wall problem. In that case, locally, it is not possible to



(a) (b) (c) (d)

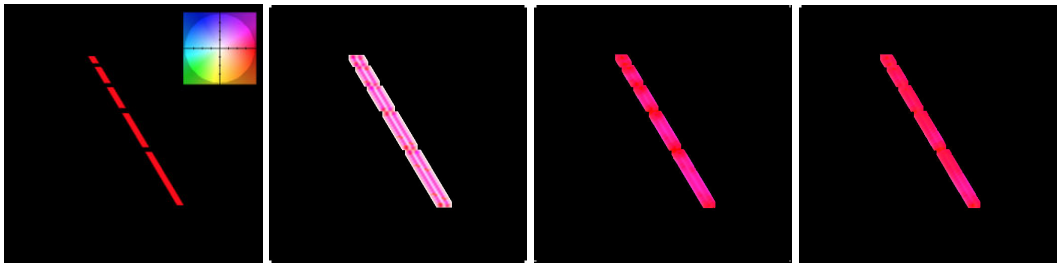
Figure 3.5: Illustration of the filling-in approach used to deal with boundary conditions and the unreliable regions. (a) How inner domain Ω_{in} (in grey) is defined taking into account V1 filter spatial size and V1 to MT pooling. Ω_{out} (in red) corresponds to $\mathcal{B}_1 \cup \mathcal{B}_2$ (see text). (b) Image domain showing the inner region Ω_{in} where exact computations can be done (i.e., without any approximation), the outer region Ω_{out} where an interpolation scheme is applied, and an example of unreliable region explained in (d). (c) Illustration of the interpolation scheme for a pixel $p \in \Omega_{out}$, showing the spatial neighbourhood associated with the spatial support of the integration and in green the region $\mathcal{A}$ which is used to estimate the interpolated values. (d) Same as (c) but in the case of an unreliable region.

find a velocity. Given a threshold T, a pixel p will be categorised as unreliable if and only if $E^{MT}(p, t, d, v^c) < T$ for all d and v^c. For these pixels, the same interpolation as (3.12) is proposed (Fig. 3.5(d)).

3.3 Results

3.3.1 Parameters settings

Table 3.2 gives parameters used in our simulations. The size of the spatial support of the V1 RF was chosen so that fine details in real-world sequences at high image resolution could be processed. V1 and MT RFs process the visual signal within an average time of 200 ms [DeAngelis 1995, Pack 2001], which corresponds to five frames for a standard video acquisition device, thus we have chosen the temporal support of the filters in order to match this constraint. With this choice, we can not have tuning to velocities higher than one pixel per frame (ppf), i.e., one ppf corresponds to the maximum temporal frequency (see (3.3)) that can be sampled for the Nyquist theorem. This limitation has been addressed here by considering a multi-scale approach, as explained in Sec. 3.2.1. The number of scales depends on the size of the input images and on the speed range (a priori unknown). For the Middlebury videos we chose six spatial scales. It is worth noting that to avoid the introduction of a loss of balance between the convolutions with the even and odd Gabor filters, the contribution of the DC component is removed [Clausi 2000]. Finally, we set the support of the spatial pooling $G_{\sigma_{pool}}$ to five which is in accordance



Ground truth	FFV1MT ($L = 1$)	FFV1MT ($L = 3$)	FFV1MT ($L = 5$)
	AAE=52.21±4.43	AAE=10.32±5.78	AAE=6.20±3.82
	EPE=1.79±0.08	EPE=0.36±0.22	EPE=0.24±0.14

Figure 3.6: Influence of the number of spatial scales. The FFV1MT model is tested with L=1, 3 and 5 scales. The color code used to show optical flow is in the inset on the first image. This color code will be used in all figures to represent optical flow. Note that the aperture problem is partially solved by considering a scale-space approach, where the effective receptive field size of MT increases and thus takes into consideration 2-D cues that are present at a distance. This can be readily observed by the results on bars with different lengths.

with findings reported in literature [Albright 1987, Bayerl 2004].

3.3.2 Analysis of proposed approaches

In this section, we evaluate the proposed FFV1MT model using synthetic and real sequences to show the intrinsic properties of our approach. When ground truth optical flow is available, average angular error (AAE) and endpoint error (EPE) will be estimated (with associated standard deviations) [Baker 2011].

The influence of the number of spatial scales is shown in Fig. 3.6. In this sequence a dashed bar moves rightward with velocity (2,0) ppf. Results show that increasing the number of scales improves the results. It is worth noting that the aperture problem is correctly solved by considering three spatial scales in the small segments, whereas five spatial scales are needed to handle longer segments, though a residual optical flow at the finest scale is not correctly recovered in the middle of the longest segment, since the spatial support of the RFs is too small with respect to the visual feature.

The next example in Fig. 3.7 is on another synthetic video that represents a textured shape moving on top of a translating background. Optical flow result show a good estimation of the optical flow except in the neighbourhood of objects boundaries (which are also here motion boundaries). The FFV1MT–TF approach looks qualitatively better, however it does not improve the quantitative performance. It might be due to the noisy texture of this synthetic sequence.

In order to analyze the roles of the different stages of the model, Fig. 3.8

Description	Parameter	Value	Equation
V1			
RF spatial scale	σ	2.27 pixels	(3.1)
... and spatial support	SS	11×11 pixels,	(3.1)
Time constant of the exp. decay	τ	2.5 frames	(3.2)
... and temporal support	TS	5 frames	(3.2)
Spatial radial peak frequency	f_s	0.25 cycles/pixel	(3.1)
Temporal radial peak frequencies	f_t	$\{0, 0.10, 0.15, 0.23\}$ cycles/frame	(3.2)
Number of spatial contrast orientations	N	8 (from 0 to π)	(5.2)
... and sampling	θ_i	$\theta = k\pi/N, k = 0..N-1$	(5.2)
Number of component speeds	M	7	(3.3)
... and sampling	v^c	$\{-0.9, -0.6, -0.4, 0, 0.4, 0.6, 0.9\}$	(3.3)
Semi-saturation constant	ε	10^{-9}	(5.2)
MT			
Std dev of the Gaussian spatial pooling	σ_{pool}	0.9 pixels	(5.1.2)
... and spatial support		5×5 pixels	(5.1.2)
Decoding step			
Number of MT direction tuning directions		2	(3.9)
... and sampling	d	$\{0, \pi/2\}$	(3.9)
Algorithm			
Number of scales	L	6	
Spatial parameter of the interpolation	α	2.5 pixels	(3.12)
Luminance parameter of interpolation	γ	1/6 of luminance range	(3.12)
Other parameters for FFV1MT-TF model			
Spatial parameter	α	$\{0.50, 0.83, 1.16, 1.50, 1.83\}$ as a function of spatial scale	(3.10)
Range parameter	β	1/6 of energy range	(3.10)
Luminance parameter	γ	1/6 of luminance range	(3.10)

Table 3.2: Parameter values used in our simulations for the FFV1MT model and its extension FFV1MT-TF. Equation number refers to the equation where it has been first introduced.

shows the V1 and MT activities. The first row shows $\|E^{V1}\|_\theta(p, v^c) = \left(\sum_{i=1}^{N} E^{V1}(p, \theta_i, v^c)^2\right)^{1/2}$: the activities do not identify specific tuning speeds, since all the spatial orientations are pooled in the norm and the tuning speeds are component speeds, i.e., they are orthogonal to the spatial orientation of the cell. The second row shows $\|E^{V1}\|_{v^c}(p, \theta) = \left(\sum_{i=1}^{M} E^{V1}(p, \theta, v_i^c)^2\right)^{1/2}$: the cells are elicited by the spatial orientation of the shape, the V1 layer shows a tuning on the spatial orientation. The third and fourth rows show $E^{MT}(p, 0, v^c)$ and $E^{MT}(p, \pi/2, v^c)$ maps, respectively. At MT layer, a speed tuning emerges: on the left, the energies are higher for the region related to the shape, this means that there is a negative speed for the horizontal and vertical velocities related to the shape. On the right, the energies are higher for the background (for the third row, only), since the background moves rightwards. These results confirm that the V1 layer has a tuning on the spatial orientation (cells respond to the spatial orientation of the shape), whereas at MT layer, a speed tuning no more related to spatial orientation emerges (i.e., the aperture problem is solved).

In Fig. 3.9 we show the distribution of E^{MT} at different positions to understand its relation to velocities. By observing the distribution of MT energies in four different positions on the original image (indicated as (a), (b), (c) and (d) in Fig. 3.7), we see how the MT layer encodes the velocities. In particular: the behaviours in (a) and (c) are affected by the values of the neighbouring borders, thus there are no prominent activities; in (b), which corresponds to a point on the foreground shape sufficiently far from borders given the actual spatial support of the filters, cells tuned to negative speeds (v_1^c) on both horizontal and vertical direction (E^{MT} with $d = 0$ and $d = \pi/2$, respectively) have the maximum response; in (d), which corresponds to a point on the background, only the response of the horizontal direction has a maximum for positive horizontal speed (v_7^c).

Fig. 3.10 shows the results of the FFV1MT model on the classical realistic Yosemite sequence with clouds. We obtain AAE=5.57 which is better than former biologically-inspired models such as the original Heeger approach (AAE=11.74, with 44.8% of reliable pixels,[Barron 1994]) and the neural model from Bayerl and Neumann (AAE=6.20, [Bayerl 2004]). One can also make comparisons with standard computer vision approaches such as Pyramidal Lucas and Kanade (AAE=6.41), modified Horn and Schunk (AAE=5.48 with 32.9% of reliable pixels, [Barron 1994]) and 3DCLG (AAE=6.18, [Bruhn 2005]), showing a better performance of the FFV1MT. The FFV1MT–TF approach shows a slightly better performance in particular close to motion discontinuities.

3.3.3 Performance evaluation on Middlebury dataset

In this section, we benchmark our approach on the computer vision dataset Middlebury [Baker 2011][1]. The sequences in this dataset bring several challenges, such as sharp edges, high velocities and occlusions. Figure 5.6 show results obtained on

[1] http://vision.middlebury.edu/flow/data/

the training dataset, which has public available ground truth. The AAEs and EPEs show that FFV1MT is able to recover reliable optical flows, though some issues remain open. Smooth effects are present on edges and fine details (see `Grove2` and `Grove3`), FFV1MT–TF partially solves this issue, as shown in `RubberWhale` and `Urban2`. The δAAE maps highlight the differences in the AAEs between FFV1MT and FFV1MT–TF, showing that the latter is better on edges as expected (red tones). In presence of high image velocity large occlusions occur, on which both approaches fail (see left-hand side of `Urban3`). In this case, the worst performance of FFV1MT–TF method is due to the fast movements of edges that undermines the luminance similarity principle on which it is based.

Figure 3.12 show results obtained on the test dataset. Higher errors coincide with occlusions (see, e.g., `Urban` sequence) and sharp edges (see, e.g., `Urban` and `Wooden` sequences), similarly to what was observed on the training set. Results can be further analysed through the Middleburg website and compared to a variety of state-of-the-art algorithms. It is worth noting that our FFV1MT model is the only neural model for motion estimation shown in the table so far.

Code

We think that this work could act as a good starting point for building scalable computer vision algorithms for motion processing that are rooted in biology. For that reason we shared the code in order to facilitate research in this direction, Matlab implementation of the FFV1MT model has been made available on ModelDB [Hines 2004]: `http://senselab.med.yale.edu/modeldb/`.

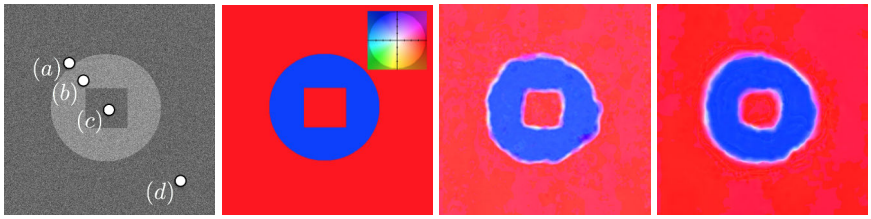
3.4 Conclusion

In this chapter, we have presented an approach that is based on a model primarily developed to account for various physiological findings related to motion processing in primates. Starting from the classical hierarchical feedforward processing model involving V1 and MT cortical areas, which is usually limited to a single spatial scale, we have extended it to consider the whole range of frequencies by adapting a multi scale approach and analysed the efficacy of the approach in estimating the dense optical flow in real world scenarios by considering an efficient velocity decoding step.

We have tested the performance of our model using synthetic stimuli as well as the standard Middlebury dataset. Results demonstrated that V1-MT feedforward model can be successfully used to compute optical flow in real videos. A qualitative evaluation indicates that model could recover velocity vectors in regions with coarse textures quite well, but typically fails to achieve robust estimates in regions with very fine texture or regions with sharp edges. This was expected, since the V1-MT feedforward model does not take into account the details of lateral interactions and scale space issues that need to be tackled in order to solve the blank wall problem and tackle regions with motion boundaries. In order to address blank wall problem, we proposed a simple extension of our baseline model using trilateral filtering at

MT level as a way to simulate lateral interactions between MT cells. Results were slightly improved suggesting that one should further focus on lateral interactions and possibly feedback into the models to better deal with real videos.

Moreover, this work has opened up several interesting questions, which could be of relevance to biologists as well, for example what could be the afferent pooling strategy of MT when there are multiple surfaces or occlusion boundaries within the MT receptive field? Can a better optical flow map be recovered by considering different multi-scale strategies? We visit these questions in the subsequent chapters.
ß



Sample input Ground truth FFV1MT FFV1MT–TF

Figure 3.7: Results on a synthetic video: A translating shape is moving with velocity $v = (-3, -3)$ ppf on top of a translating background moving with velocity $v = (4, 0)$ ppf. Results are AAE=3.56±14.40, EPE=0.26±0.86. for FFV1MT and AAE=3.70±14.78, EPE=0.27±0.86 for FFV1MT–TF.

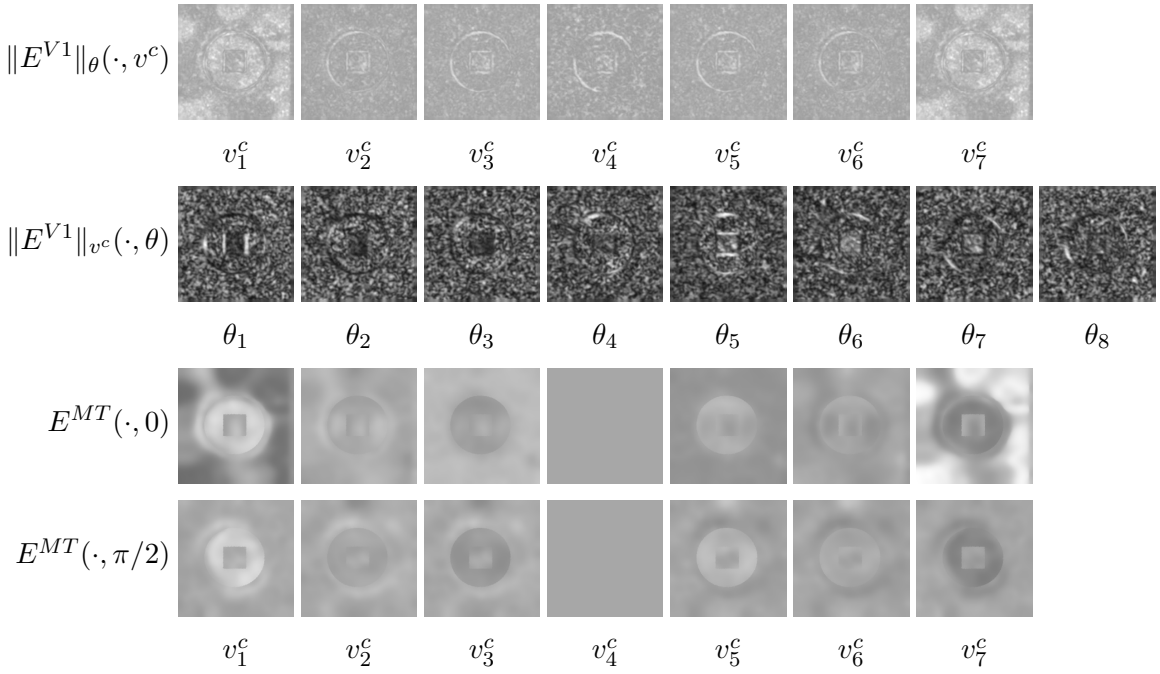


Figure 3.8: V1 and MT activities on the synthetic video shown in Fig. 3.7 (see text).

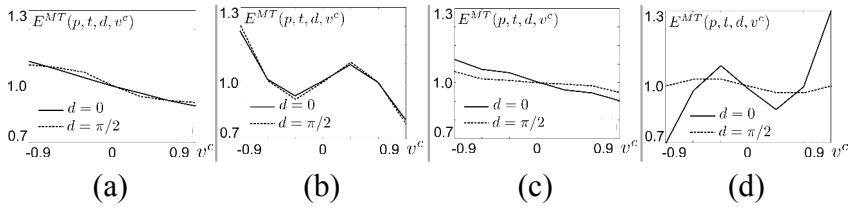
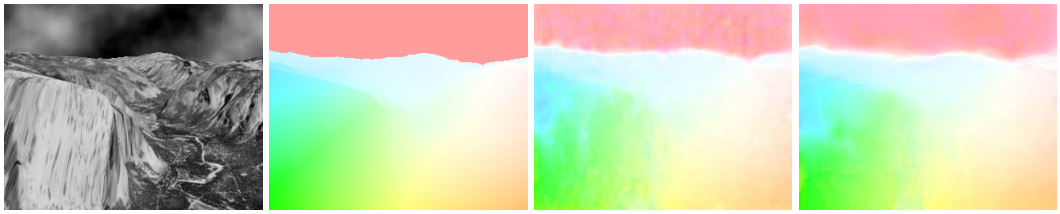
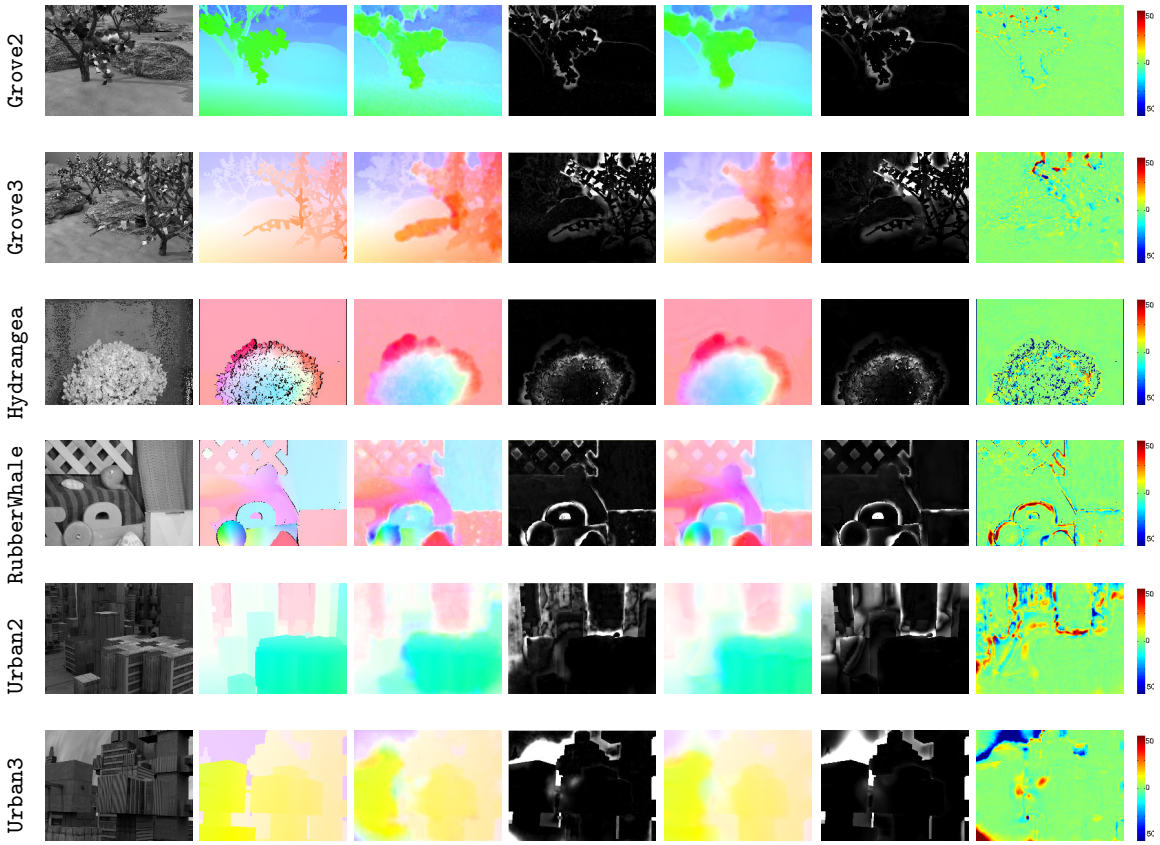


Figure 3.9: Distribution of MT energy at positions indicated in Fig. 3.7.



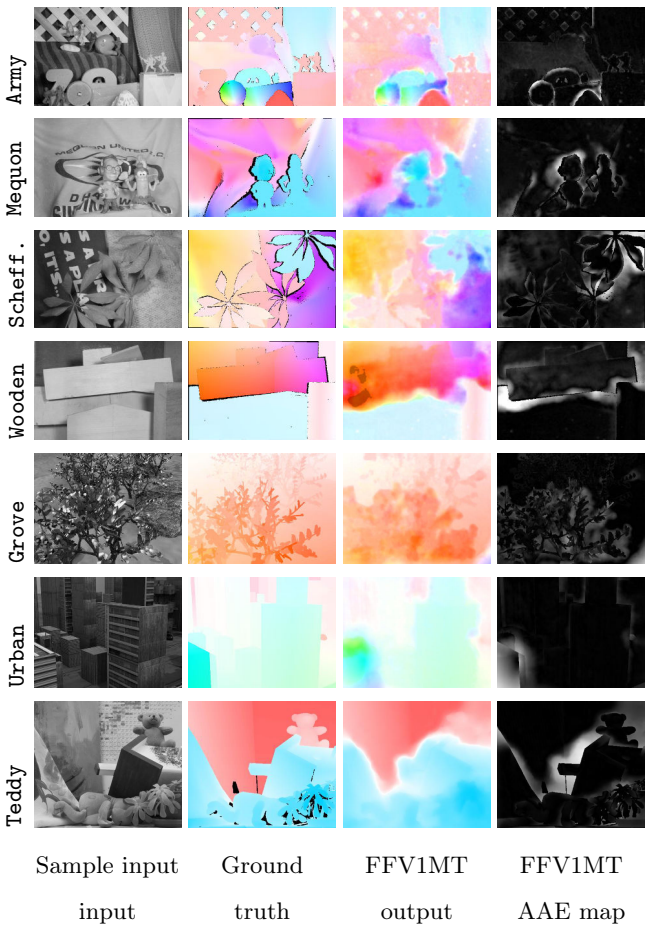
Sample input	Ground truth	FFV1MT	FFV1MT–TF
		AAE=5.57 ± 12.27	AAE=4.58 ± 9.35
		EPE=0.23 ± 0.39	EPE=0.20 ± 0.33

Figure 3.10: Performance of the FFV1MT and FFV1MT–TF models on the classical `Yosemite` sequence with clouds. The color code is the same as in Fig. 3.6.



Sequence	FFV1MT		FFV1MT–TF	
	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD
grove2	4.28 ± 10.25	0.29 ± 0.62	3.96 ± 10.00	0.26 ± 0.60
grove3	9.72 ± 19.34	1.13 ± 1.85	9.40 ± 18.39	1.10 ± 1.74
Hydrangea	5.96 ± 11.17	0.62 ± 0.96	5.74 ± 11.03	0.59 ± 0.86
RubberWhale	10.20 ± 17.67	0.34 ± 0.54	9.38 ± 15.21	0.31 ±0.46
urban2	14.51 ± 21.02	1.46 ± 2.13	13.62 ± 10.23	1.19 ±1.58
urban3	15.11 ± 35.28	1.88 ± 3.27	16.22 ± 2.21	1.69 ±2.85

Figure 3.11: Sample results and error measurements on Middlebury training set. δAAE=$AAE_{FFV1MT} - AAE_{FFV1MT--TF}$ is represented with a color code, where red and blue tones are for positive and negative values, respectively.



Sequence	AAE All (Rank) - Disc. (Rank)	EPE All (Rank) - Disc. (Rank)
Army	12.02(102) - 23.3(102)	0.33(100) - 0.64(100)
Mequon	10.7(94) - 26.6(103)	0.79(94) - 1.90(103)
Schefflera	15.6(96) - 29.0(101)	1.33(104) - 1.90(101)
Wooden	16.6(102) - 36.3(105)	1.38(103) - 2.98(104)
Grove	6.51(105) - 6.40(103)	1.76(105) - 1.99(105)
Urban	16.2(104) - 30.7(105)	2.33(105) - 3.64(106)
Yosemite	3.41(74) - 5.44(88)	0.16(66) - 0.18(83)
Teddy	12.3(101) - 18.8(102)	1.81(100) - 2.64(100)

Figure 3.12: Sample results and error measurements of FFV1MT model on Middlebury test set. By the time of evaluation 107 algorithms are benchmarked by the website, and Rank indicates the relative performance of the method with respect to others for both the entire sequence (All) and for discontinuities (Disc.). The results are public at http://vision.middlebury.edu/flow/eval

Chapter 4

Adaptive pooling and activity spread for Optical Flow

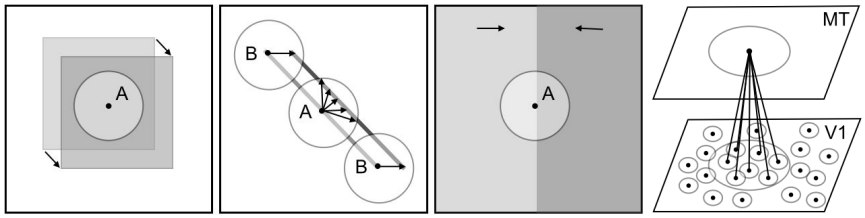
In this chapter, we study the impact of local context of an image (contrast and 2D structure) on spatial motion integration by MT neurons. This study has been inspired by the limitations of the feedforward V1-MT model presented in the previous chapter to deal with the complex real world scenes. In particular, we address the difficulties encountered by the model due to isotropic spatial pooling in regions close to texture/motion boundaries. The key contributions of this chapter can be stated as follows:

- We propose an extension to the FFV1MT model with adaptive processing by focussing on the role of local context which indicates the reliability of the local velocity estimates. The extended model takes into consideration a network structure representative of V1, V2 and MT areas.
- We incorporate three functional principles observed in primate visual system: contrast adaptation, adaptive afferent pooling and MT activity spread that is conditioned upon the 2D image structure.
- We evaluate the proposed model, referred to as APMD (Adaptive Pooling and Motion Diffusion) using the Middlebury optical flow estimation dataset. Results demonstrated improved performance by the AMPD model compared to the baseline FFV1MT model.

4.1 Local motion analysis by FFV1MT model

For each neuron in the cortical hierarchy, one can associate a receptive field defined by the region in the visual field that elicits a response. Receptive fields are first small and become larger going deeper in the hierarchy [Orban 2008]. The first local analysis of motion is done at the V1 cortical level. The small receptive field size of V1 neurons, and their strong orientation selectivity, poses several difficulties when estimating global motion direction and speed, as explained in Sec. 2.1 and illustrated in Fig. 4.1. In particular, any local motion analyzer will face blankwall problem, aperture problem and selection ambiguity due to presence of multiple motions [Bradley 2008].

In terms of optical flow estimation, feedforward computation involving V1 and MT could be sufficient in the case of regions without any ambiguity. On the contrary,



(a) (b) (c) (d)

Figure 4.1: Estimation motion from local observations: (a) blank wall problem: at position A, the absence of texture gives no information to estimate motion; (b) aperture problem: at position A, only the 1D component of the flow is known; (c) multiple motion: at position A, receptive field integrates different motion informations; (d) Illustration of a pooling step with the corresponding receptive fields.

recovering velocity at regions where there is local ambiguity such as the aperture or the blank wall problems would require pooling reliable information from far, less ambiguous regions in the surrounding. Such non-local information is thought to be conveyed by the intricate network interactions (short-range, or recurrent networks, and long-range) often involving areas processing form based cues such as V2 and V4 (see [Masson 2010] for reviews).

4.2 Biological vision solution

4.2.1 Cortical hierarchy

Here, we present a caricature of the motion processing pipeline that has been suggested after extensive studies in monkeys [Orban 2008], an extension to the minimalistic feedforward V1-MT view. The cortical areas considered V1, V2 and MT are illustrated in Figure 4.2. Motion information is extracted locally through a set of spatiotemporal filters tilling the retinotopic representation of the visual field in area V1. However, these direction-selective cells exhibit several non-linear properties as the center response is constantly modulated by the surrounding inputs conveyed by feedback and lateral inputs. Context modulations are not only implemented by center-surround interactions in areas V1 and MT. For instance, other extra-striate areas such as V2 or V4 project to MT neurons to convey information about the structure of the visual scene, such as the orientation or color of local edges.

4.2.2 Contrast adaptive processing

The structure of neuronal receptive fields is not static as it has long been thought [Fairhall 2014]. Rather, it adapts to the local context of the image so that many of the tuning functions characterizing low-level neurons are in fact dynamical (e.g. [Sharpee 2006]). A first series of evidence comes from experiments where the properties of the local inputs change the classical receptive field. For

instance orientation-tuning in area V1 and speed tuning of MT neurons are sharper when tested with broad-band texture inputs, as compared to low-dimension gratings (e.g., [Freeman 2013, Priebe 2003]). Moreover, spatial summation function often broadens as contrast decreases or noise level increases [Sceniak 1999]. These observations are complemented by experiments varying the spatial context of this local input. For instance, surround inhibition in V1 and MT neurons becomes stronger at high contrast and center-surround interactions exhibit a large diversity in terms of their relative tunings. Moreover, the spatial structure of these interactions is often more diverse in shape than the classical Mexican-hat (see [Bradley 2008] for a review). Lastly, at each decoding stage, it seems nowadays that tuning functions are weighted by the reliability of the neuronal responses, as varying for instance with contrast or noise levels [Ma 2014]. Still, these highly adaptive properties have barely been taken into account when modelling visual motion processing. Here, we model some of these mechanisms to highlight their potential impact on optic flow computation. We focus on both the role of local image structure (contrast, textureness) and the reliability of these local measurements in controlling the spatial propagation mechanisms. We investigated how these mechanisms can help solving local ambiguities, and segmenting the flow fields into different surfaces while still preserving the sharpness and precision of flow estimation.

4.3 Extended Model (APMD)

The baseline model involving a feedforward processing from V1 to MT is largely devised to describe physiological and psychophysical observations on motion estimation when the testing stimuli were largely homogeneously textured regions such as moving gratings and plaids. Hence the model is limited in the context of dense

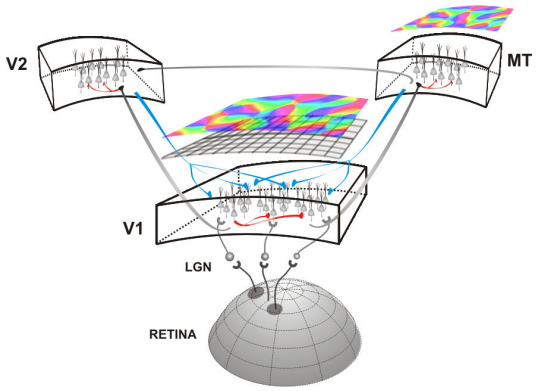


Figure 4.2: Illustration of the motion processing pathway, with the main cortical areas involved in motion estimation (area V1) and integration (area MT). Interactions with the form pathway are represented by V2 and V4 cortical areas. This cartoon illustrates the variety of connectivities: feedforward (in gray), short- and long-range lateral (in red) and feedback (in blue).

flow estimation for natural videos as it has no inherent mechanism to deal with associated sub problems such blank wall problem, aperture problem or occlusion boundaries.

Building on recent results summarized in Sec. 4.2.2 we model some of these mechanisms to highlight their potential impact on optic flow computation. Considering inputs from area V2, we focus on the role of local context (contrast and image structure) indicative of the reliability of these local measurements in (i) controlling the pooling from V1 to MT and (ii) adding lateral connectivity in MT.

4.3.1 Area V2: Contrast and Image Structure

Our goal is to define a measure of contrast which is indicative of the aperture and blank wall problems using the responses of spatial Gabor filters. There exist several approaches to characterize the spatial content of an image from Gabor filter. For example, in [Kovesi 1999] the authors propose the phase congruency approach which detects edges and corners irrespectively of contrast in an image. In dense optical flow estimation problem, region with texture are less likely to suffer blank wall and aperture problems even though edges are susceptible to aperture problem. So phase congruency approach cannot be used directly and we propose the following simple alternative approach.

Let h_{θ_i} the Gabor filter for edge orientation θ_i, we define

$$R(p) = (R_{\theta_1}(p), \ldots, R_{\theta_N}(p)) \text{ where } R_{\theta_i}(p) = |h_{\theta_i} * I|(p).$$

Given an edge orientation at θ_i, R_{θ_i} is maximal when crossing the edge and ∇R_{θ_i} indicate the direction to go away from edge.

Then the following contrast/cornerness measure is proposed as follows, taking into consideration the amount of contrast at a given location and also ensuring that contrast is not limited to a single orientation giving raise to aperture problem.

$$\mu(R)(p) = \frac{1}{N} \sum_i R_{\theta_i}(p), \tag{4.1}$$

$$C(p) = H_\xi(\mu(R(p))(1 - \sigma^2(R(p))/\sigma^2_{max}), \tag{4.2}$$

where $\mu(R(p))$ (resp. $\sigma^2(R(p))$) denote the average (resp. variance) of components of R at position p, $H_\xi(s)$ is a step function ($H_\xi(s) = 0$ if $s \leq \xi$ and 1 otherwise) and $\sigma^2_{max} = \max_{p'} \sigma^2(R(p))$. The term $H_\xi(\mu(R(p))$ is an indicator of contrast as it measures the Gabor energies: in regions with strong contrast or strong texture in any orientation this term equals to one; in a blank wall situation, it is equal to zero. The term $(1 - \sigma^2(R(p))/\sigma^2_{max})$ measures how strongly the contrast is oriented in a single direction: it is higher when there is only contrast in one direction and lower when there is contrast in more than one orientation (thus it is an indicator of where there is aperture problem).

4.3.2 Area MT: V2-Modulated Pooling

Most of the models currently pool V1-afferents using a linear fixed receptive field size, which does not adapt itself to the local gradient or respect discontinuities in spatio-temporal reposes. This might lead to degradation in the velocity estimates by blurring edges/kinetic boundaries. Thus it is advantageous to make the V1 to MT pooling adaptive as a function of texture edges.

We propose to modify the pooling stage as follows

$$E^{MT}(p,t;d,v^c) = F\left(\sum_{i=1}^{N} w_d(\theta_i)\tilde{\mathcal{P}}(E^{V1})(p,t;\theta_i,v^c)\right),$$

where the spatial pooling become functions of image structure. We propose the following texture-dependent spatial pooling:

$$\mathcal{P}(E^{V1})(p,t;\theta_i,v^c) = \frac{1}{\bar{N}(p,\theta_i)}\sum_{p'} f_{\alpha(\|R\|(p))}(\|p-p'\|)g_i(p,p')E^{V1}(p,t;\theta_i,v^c) \tag{4.3}$$

where $\bar{N}(p,\theta_i) = \sum_{p'} f_{\alpha(\|R\|(p))}(\|p-p'\|)g_i(p,p')$ is a normalizing term. The two weights are now depending on image structure. The variance of the distance term α, i.e., the size of the integration domain, now depends on the structure R_{θ_i} as follows

$$\alpha(\|R\|(p)) = \alpha_{max}e^{-\eta\frac{\|R\|^2(p)}{r_{max}}}, \tag{4.4}$$

where η is a constant, $r_{max} = \max_{p'}\{\|R\|^2(p')\}$. Then there is an additional term $g_i(p,p')$ to enable anisotropic pooling close to image structures so that discontinuities could be better preserved. Here we propose to define g_i by

$$g_i(p,p') = S_{\lambda,\nu}\left(-\frac{\nabla R_{\theta_i}(p)}{\|\nabla R_{\theta_i}\|+\varepsilon}\cdot(p'-p)\right), \tag{4.5}$$

where $S_{\lambda,\nu} = 1/(1+\exp(-\lambda(x-\nu)))$ is a sigmoid function and ε a small constant. Note that this term is used only in regions where $\|\nabla R_{\theta_i}\|$ is greater than a threshold. Fig. 4.3(a) gives examples of the pooling coefficients at different p locations.

4.3.3 MT Lateral Interactions

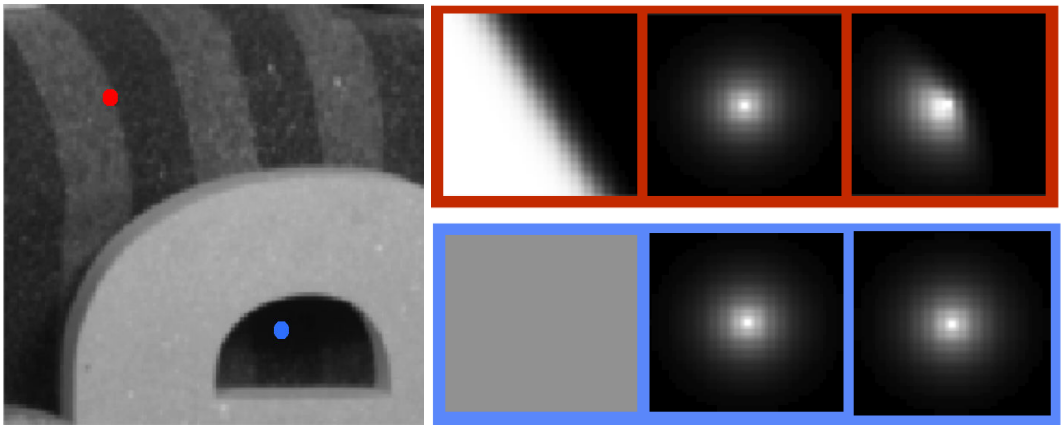
We model the lateral iterations for the velocity information spread (from the regions where there is less ambiguity to regions with high ambiguity, see Sec. 4.1) whilst preserving discontinuities in motion and illumination. To do so, we propose an iterated trilateral filtering defined by:

$$u^{n+1}(p) = \frac{1}{\bar{N}(p)}\sum_{p'} W(p,p')u^n(p'), \tag{4.6}$$

$$c^{n+1}(p) = c^n(p) + \lambda(\max_{p'\in\mathcal{N}(p)} c^n(p') - c^n(p)) \tag{4.7}$$

$$u^0(p) = E^{MT}(p,t;\theta_i,v^c), \tag{4.8}$$

$$c^0(p) = C(p), \tag{4.9}$$



(a) (b) (c) (d)

Figure 4.3: Impact of local contrast on pooling strategy: (a) Sample input indicating two different locations being sampled. (b) (A) Anisotropy term, (B) Spatial term, (C) Pooling region.

where

$$W(p,p') = c^n(p') f_\alpha(\|p-p'\|) f_\beta(c^n(p)(u^n(p') - u^n(p))) \\ f_\gamma(I(p') - I(p)) u^n(p'), \quad (4.10)$$

and $\mathcal{N}(p)$ is a local neighbourhood around p. The term $c(p')$ ensures that more weight is given naturally to high confidence estimates; The term $c(p)$ inside f_β ensures that differences in the MT responses are ignored when confidence is low facilitating the diffusion of information from regions with high confidence and at the same time preserves motion discontinuities or blurring at the regions with high confidence.

4.4 Results

In order to test the method a multi-scale version of both the baseline approach (FFV1MT) and approach with adaptive pooling (APMD) are considered. The method is applied on a Gaussian pyramid with 6 scales, the maximum number of scales that could be reliably used for the spatio-temporal filter support that has been chosen.

A first test was done on Yosemite sequence as it is widely used in both computer vision and biological vision studies (see Fig. 4.4, first row). For APMD method we obtain AAE=3.00 $\pm$ 2.21. This can be compared to what has been obtained with previous biologically-inspired models such as the original Heeger approach (AAE=11.74°, but estimated 44.8% of the most reliable regions, see [Barron 1994]) and the neural model from Bayerl and Neumann (AAE=6.20°, [Bayerl 2004]), showing an improvement. One can do comparisons with standard computer vision ap-

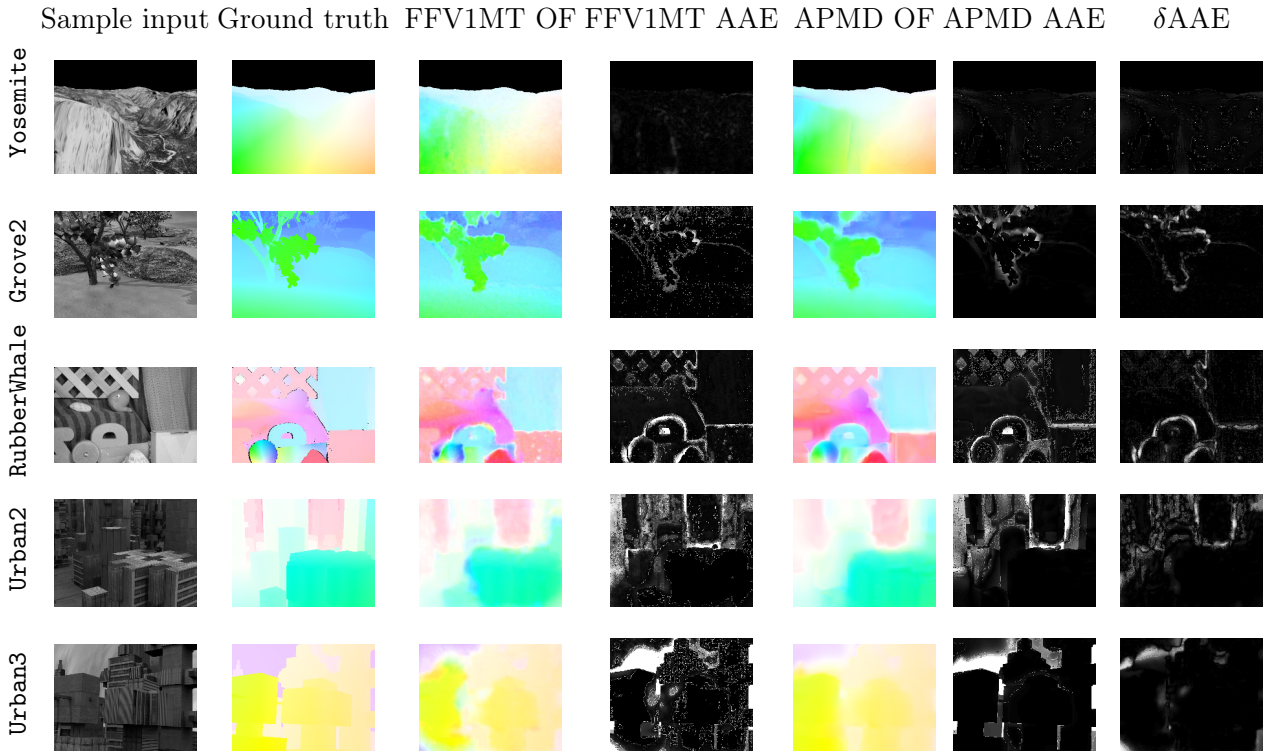


Figure 4.4: Sample results on Yosemite sequence and a subset of Middlebury training set. δAAE = AAE_{FFV1MT} - AAE_{APMD}

proaches such as Pyramidal Lucas and Kanade (AAE=6.41°) and Horn and Schunk (AAE=6.41°, [Horn 1981]), showing comparable performance.

The results on the Middlebury training set show improvements of the proposed method with respect to the FFV1MT(see Table 4.1). For qualitative comparison, sample results are also presented in Fig.4.4. The relative performance of extended method can be understood by observing δAAE, difference between the FFV1MT AAE map and the AAE map of APMD which are presented in Fig. 4.4 (last column): the improvements are prominent at the edges, e.g. see the δAAE column for the RubberWhale and Urban2 sequence. A close up view of the results obtained for RubberWhale sequence are presented in Fig. 4.5. Qualitatively the sharp details of the object boundaries in the motion map are much better preserved in case of APMD when compared to FFV1MT. Th

This model still has several limitations:

- Noisy edge extraction: The spatio-temporal filter based edge extraction is still noisy. At lower scales if the pooling direction is wrongly estimated it would effect the flow estimation for considerably large region as we rely on the warping scheme.

- Limitations of the anisotropy: The anisotropic pooling that has been consid-

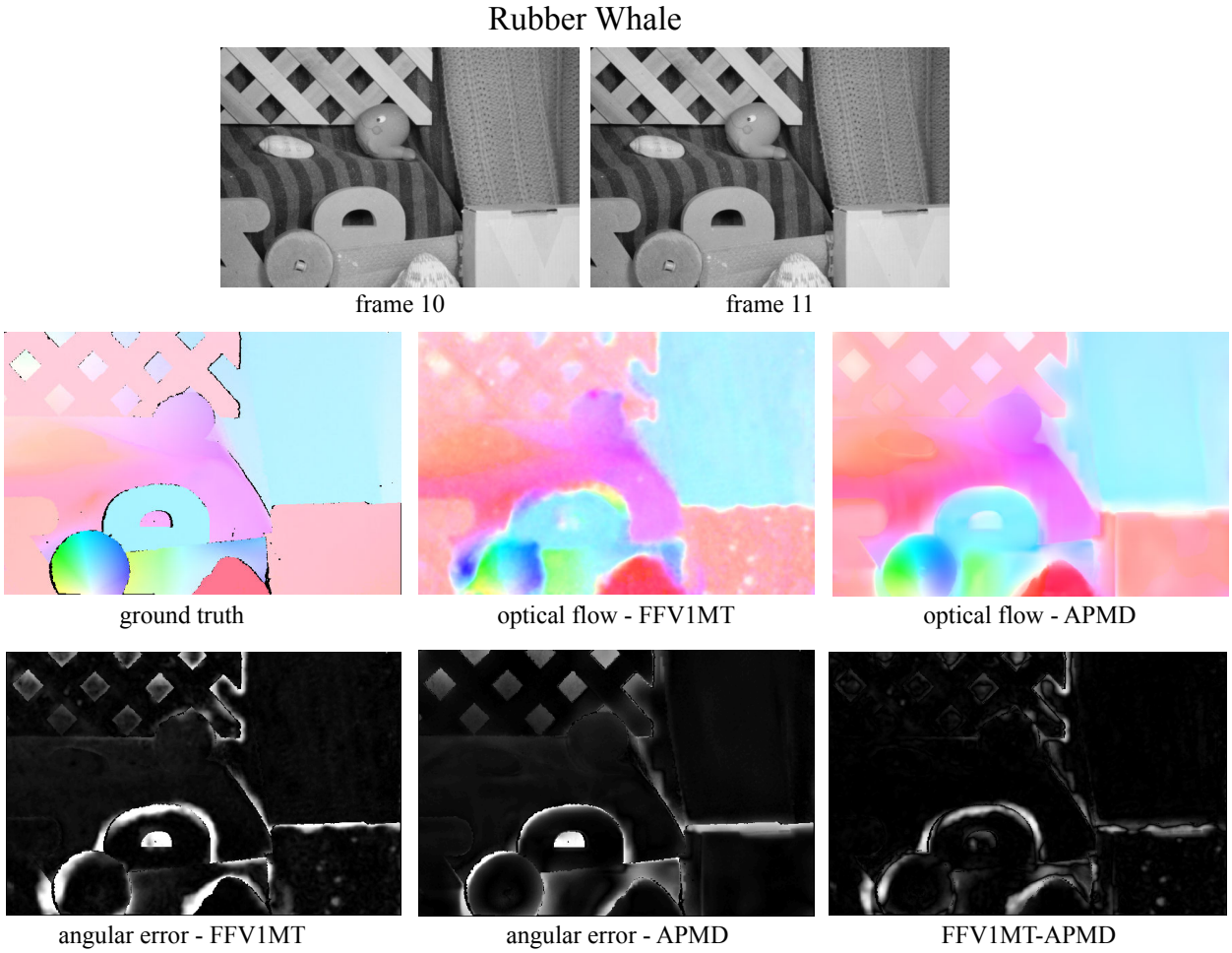


Figure 4.5: Results on rubber whale sequence from Middlebury training dataset.

	FFV1MT		APMD	
Sequence	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD
grove2	4.28 ± 10.25	0.29 ± 0.62	4.07 ± 9.29	0.27 ± 0.56
grove3	9.72 ± 19.34	1.13 ± 1.85	10.66 ± 19.25	1.11 ± 1.61
Hydrangea	5.96 ± 11.17	0.62 ± 0.96	5.48 ± 11.10	0.50 ± 0.69
RubberWhale	10.20 ± 17.67	0.34 ± 0.54	8.87 ± 13.16	0.30 ± 0.42
urban2	14.51 ± 21.02	1.46 ± 2.13	12.70 ± 19.92	1.09 ± 1.31
urban3	15.11 ± 35.28	1.88 ± 3.27	12.78 ± 31.36	1.32 ± 2.25

Table 4.1: Error measurements on Middlebury training set

ered does not always ensure that signals from different motion surfaces are not combined together. Particularly in regions which have very fine structures.

- Lack of inter-scale interactions: The model is originally inspired by the proposal from Heeger et al. [Heeger 1988, Simoncelli 1998], which analyses the filter responses at multiple scales simultaneously and thus could be picking up on the best possible scale for analysis. In our model we don't have inter-scale interactions.

- MT decoding: Currently the model uses linear decoding scheme by sparsely sampling the velocity along the cardinal axes. This would mean that at motion boundaries instead of selecting one of the motion component the model's estimated velocity is shifted towards the other component due to averaging.

4.5 Conclusion

In this chapter, we have proposed a new algorithm that incorporates three functional principles observed in primate visual system, namely contrast adaptation, image structure based afferent pooling and ambiguity based lateral interaction. Even though, local context based information propagation and adaptive pooling seem to improve the flow estimation near the object/motion boundaries, the model does not outperform the latest computer vision methods. One particular aspect is that, by considering a scale-space and warping approach, we are not taking into account inter-scale interactions and by also considering a linear decoding scheme we are missing out on over-all population shape that might be indicative of the presence of multiple motion surfaces, such as a transparent surface/motion boundary. In the next chapter, we examine the velocity decoding problem and also the impact of considering the spatio-temporal filter outputs at multiple scales simultaneously.

Chapter 5

Decoding MT Motion Response for Optical Flow Estimation

Representation of motion in terms of spatio-temporal motion energies extracted at V1 stage remains a dominant hypothesis in visual neuroscience. Thus, decoding the motion energies is of natural interest for developing biologically inspired computer vision algorithms for dense optical flow estimation. In this chapter:

- We address the decoding problem by evaluating four strategies for motion estimation starting from spatio-temporal energies extracted at V1 stage: intersection of constraints, maximum likelihood, linear regression on MT responses and neural network based regression using multi scale-features.
- We characterize the performances and the current limitations of the different strategies, in terms of recovering dense flow estimation using Middlebury benchmark dataset widely used in computer vision and highlight key aspects for future developments.

The chapter is organised as follows. In Sect. 5.1, we present the basis of this approach which is a feedforward model of V1 and MT cortical areas response. It is a summary of the model presented in chapter 3. In Sect. 5.2, given MT population responses, we propose four decoding strategies to estimate optical flow. These four strategies are then evaluated and discussed in Sect. 5.3 using classical sequences from the literature.

5.1 V1-MT model for motion processing

5.1.1 Area V1: Motion Energy

Let us consider a gray scale image sequence $I(p,t)$, for all positions $p = (x,y)$ inside a domain Ω and for all time $t > 0$. Our goal is to find the optical flow $v(p,t) = (v_x, v_y)(p,t)$ defined as the apparent motion at each position p and time t.

Simple cells are characterized by the preferred spatial orientation θ of their contrast sensitivity in the spatial domain and their preferred velocity v^c in the direction orthogonal to their contrast orientation often referred to as component speed. The receptive fields of the V1 simple cells are classically modelled using band-pass filters in the spatio-temporal domain. In order to achieve low computational complexity, the spatio-temporal filters are decomposed into separable filters in space and

time. Spatial component of the filter is described by Gabor filters h and temporal component by an exponential decay function k. We define the following complex filters:

$$h(p;\theta,f_s) = Be^{\left(\frac{-(x^2+y^2)}{2\sigma^2}\right)}e^{j2\pi(f_s cos(\theta)x + f_s sin(\theta)y)},$$
$$k(t;f_t) = e^{\left(-\frac{t}{\tau}\right)}e^{j2\pi(f_t t)},$$

where σ and τ are the spatial and temporal scales respectively, which are related to the spatial and temporal frequencies f_s and f_t and to the bandwidth of the filter. Denoting the real and imaginary components of the complex filters h and k as h_e, k_e and h_o, k_o respectively, and a preferred velocity (speed magnitude) $v_c = f_t/f_s$, we introduce the odd and even spatio-temporal filters defined as follows,

$$g_o(p,t;\theta,v^c,\sigma) = h_o(p;\theta,f_s)k_e(t;f_t) + h_e(p;\theta,f_s)k_o(t;f_t),$$
$$g_e(p,t;\theta,v^c,\sigma) = h_e(p;\theta,f_s)k_e(t;f_t) - h_o(p;\theta,f_s)k_o(t;f_t).$$

These odd and even symmetric and tilted (in space-time domain) filters characterize V1 simple cells. Using these expressions, we define the response of simple cells, either odd or even, with a preferred direction of contrast sensitivity θ in the spatial domain, with a preferred velocity v^c and with a spatial scale σ by

$$R_{o/e}(p,t;\theta,v^c,\sigma) = g_{o/e}(p,t;\theta,v^c,\sigma) \overset{(p,t)}{*} I(p,t). \tag{5.1}$$

The complex cells are described as a combination of the quadrature pair of simple cells (5.1) by using the motion energy formulation,

$$E(p,t;\theta,v^c,\sigma) = R_o(p,t;\theta,v^c,\sigma)^2 + R_e(p,t;\theta,v^c,\sigma)^2,$$

followed by a normalization. Assuming that we consider a finite set of orientations $\theta = \theta_1 \ldots \theta_N$, the final V1 response is given by

$$E^{V1}(p,t;\theta,v^c,\sigma) = \frac{E(p,t;\theta,v^c,\sigma)}{\sum_{i=1}^{N} E(p,t;\theta_i,v^c,\sigma) + \varepsilon}, \tag{5.2}$$

where $0 < \varepsilon \ll 1$ is a small constant to avoid divisions by zero in regions with no energies, which happens when no spatio-temporal texture is present.

5.1.2 Area MT: Pattern Cells Response

MT neurons exhibit velocity tuning irrespective of the contrast orientation. This is believed to be achieved by pooling afferent V1 responses in both spatial and orientation domains followed by a non-linearity [Simoncelli 1998]. The response of a MT pattern cell tuned to the speed v^c and to direction of speed d can be expressed as follows:

$$E^{MT}(p,t;d,v^c,\sigma) = F\left(\sum_{i=1}^{N} w_d(\theta_i)\mathcal{P}(E^{V1})(p,t;\theta_i,v^c,\sigma)\right),$$

where w_d represents the MT linear weights that give origin to the MT tuning (see example in Fig. 5.1). It can be defined by a cosine function shifted over various orientations [Maunsell 1983a, Rust 2006], i.e.,

$$w_d(\theta) = \cos(d - \theta) \quad d \in [0, 2\pi[.$$

Then, $\mathcal{P}(E^{V1})$ corresponds to the spatial pooling and is defined by

$$\mathcal{P}(E^{V1})(p, t; \theta_i, v^c, \sigma) = \frac{1}{A} \sum_{p'} f_\alpha(\|p - p'\|) E^{V1}(p, t; \theta_i, v^c, \sigma), \tag{5.3}$$

where $f_\alpha(s) = \exp(s^2/2\alpha^2)$, $\|.\|$ is the L_2-norm, α is a constant, A is a normalization term (here equal to $2\pi\alpha^2$) and $F(s) = exp(s)$ is a static nonlinearity chosen as an exponential function [Rust 2006]. The pooling defined by (5.3) is a spatial Gaussian pooling.

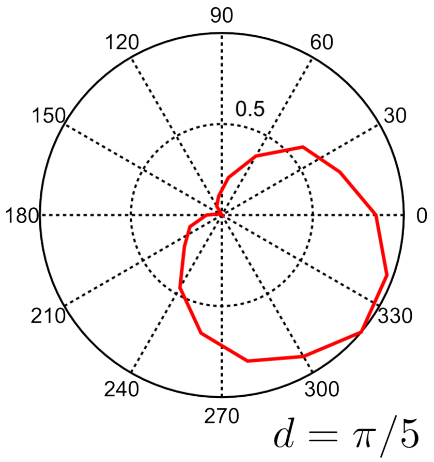


Figure 5.1: Example of a MT direction ($d = \pi/5$) tuning curve for moving plaid stimuli that span all the speed directions.

5.2 Decoding of the velocity representation of area MT

In order to engineer an algorithm capable of recovering dense optical flow estimates, we need to address the problem of decoding the population responses of tuned MT neurons. Indeed, a unique velocity vector cannot be recovered from the activity of a single velocity tuned MT neuron as multiple scenarios could evoke the same activity. However, a unique vector can be recovered from the population activity. In this paper, the velocity space was sampled by considering MT neurons that span over the 2-D velocity space with a preferred set of tuning speed directions d in $[0, 2\pi[$ and also a multiplicity of tuning speeds v^c.

Four strategies are described below. The first three strategies, called *intersection of constraints*, *maximum likelihood* and *learned linear* decoding are based on coarse-to-fine approach in order to consider multiple spatial frequencies f_s and to compute

large velocities. This approach is illustrated in Fig. 5.2 and described in chapter 3. Here the decoding stage will impact the quality of the optical flow extracted at each scale as it is used for the warping. In the third strategy an optimal linear decoding is learned and is applied at each scale. Alternatively, the fourth strategy, called *regression with neural network*, learns to estimate optical flow by considering the V1 responses at all the scales together.

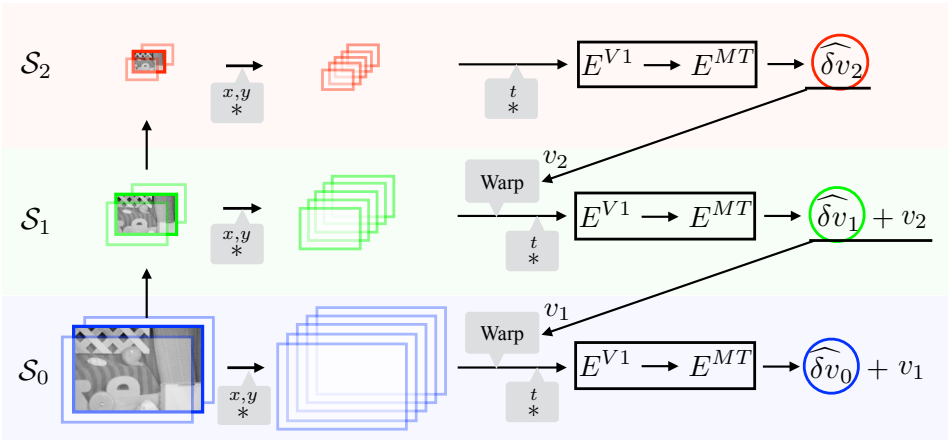


Figure 5.2: Coarse-to-fine approach for optical flow based on a V1-MT model 3. At each scale, decoding is needed to warp V1 motion energies at the coarser scale.

5.2.1 Intersection of Constraints Decoding

The MT responses are obtained through a static nonlinearity described by an exponential function, thus we can linearly decode the population activities [Rad 2011]. Since the distributed representation of velocity is described as a function of two parameters (speed and direction), first we linearly decode the speed (velocity magnitude) for each speed direction, then we apply the intersection of constraints (IOC) mechanism [Bradley 2008] to compute the speed direction. The speed along direction d can be expressed as:

$$v^d(p,t;d,\sigma) = \sum_{v_i^c=v_1^c}^{v_M^c} v_i^c E^{MT}(p,t;d,v_i^c,\sigma). \tag{5.4}$$

Then the IOC solution is defined by:

$$\begin{aligned} &\vec{v} = \underset{\vec{w}}{\operatorname{argmin}} \{G(\vec{w})\}, \\ &\text{where } G(\vec{v}) = \textstyle\sum_{d_i=d_1}^{d_Q} (v^{d_i} - \vec{v} \cdot [\cos d_i \sin d_i]^T)^2, \end{aligned} \tag{5.5}$$

where $(\cdot)^T$ indicates the transpose operation. The analytic solution of Eq. 5.5 gives:

$$\begin{aligned} v_x &= \frac{2}{Q}\sum_{d_i=d_1}^{d_Q} v^d(p,t;d_i,\sigma)\cos d_i \\ v_y &= \frac{2}{Q}\sum_{d_i=d_1}^{d_Q} v^d(p,t;d_i,\sigma)\sin d_i. \end{aligned} \tag{5.6}$$

5.2.2 Maximum Likelihood Decoding

The MT activities (see Fig 5.3 for an example of a MT population response that shows a peak for the direction and the speed present in the input stimulus) can be decoded with a Maximum Likelihood (ML) technique [Pouget 1998]. In this

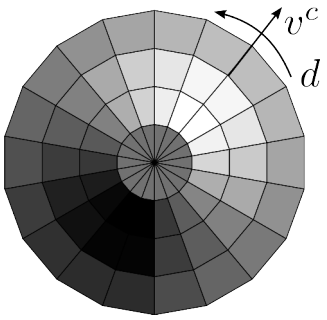


Figure 5.3: An example of MT population response at a given image point p, for a random dot sequence that moves at $v_x = 0.3$ and $v_y = 0.3$ pixel/frame. The speed directions d have 19 values in the range $[0, 2\pi[$, and the speeds v^c have 7 values in the range ± 1 pixel/frame.

paper, the ML estimate is performed through a curve fitting, or template matching, method. In particular, we decode the MT activities by finding the Gaussian function that best match the population response. The position of the peak of the Gaussian corresponds to the ML estimate.

5.2.3 Linear Decoding Through Learned Weights

We can learn the two-dimensional matrix of weights $\mathcal{W}$ that are used to linearly decode the MT activities ($\vec{v} = E^{MT}\mathcal{W}$ for each image pixel p). To learn such weights, we have considered a dataset of 8×7 random dot sequences with known speeds (both v_x and v_y, 8 directions and 7 speeds), which cover the spatio-temporal filters' range, and we have minimized a cost function to compute the best weights $\mathcal{W}$. The cost function is defined by:

$$||\mathcal{R}\mathcal{W} - v_{gt}||^2 + \lambda||\mathcal{W}||^2, \tag{5.7}$$

where $\mathcal{R}$ is a matrix whose rows contain the MT population responses (for the whole training set), $\mathcal{W}$ is the vector of weights, and v_{gt} contains the ground truth speeds. It is worth to note that such procedure has been carried out at a single spatial scale. Since we use random dots, we have considered the average MT responses, and $\lambda = 0.05$. Figure 5.4 shows the learned two-dimension matrix of weights.

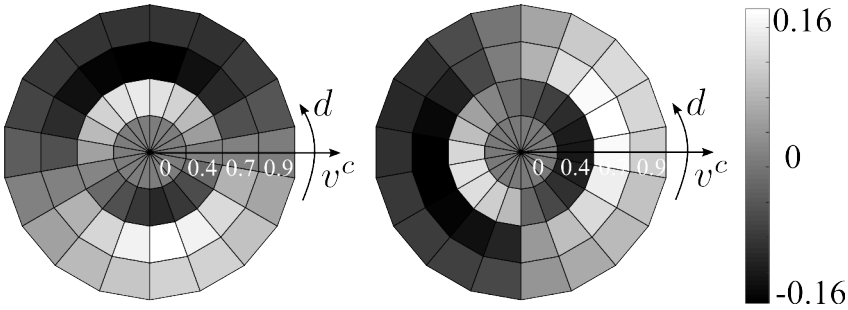


Figure 5.4: Two-dimensional matrix of weights learned through sequences of random dots. The matrices on the left and on the right are used to decode v_x and v_y, respectively.

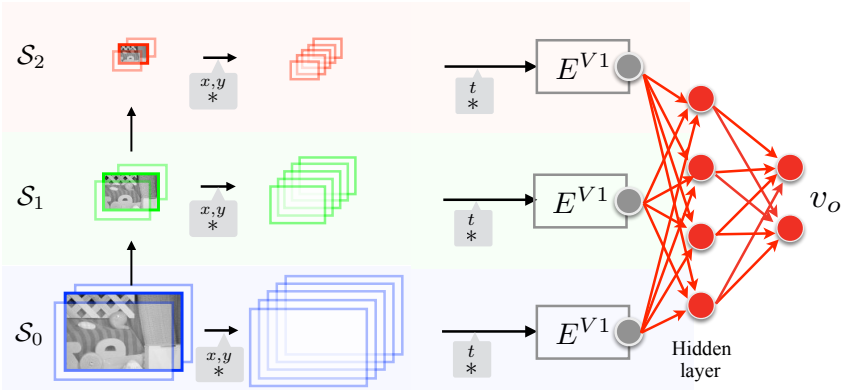


Figure 5.5: Neural network based regression for optical flow estimation.

5.2.4 Decoding with Regression using Neural Network

For the regression using neural network, spatio-temporal energies representative of the V1 complex cell responses are computed across various scales and are concatenated to form an input vector of dimension 504 (6 scales × 12 orientations × 7 velocities). The feature computation stage is illustrated in Fig. 5.5. It is worth to note that in this decoding strategy we do not use the coarse to fine approach. A feedforward network comprising of a hidden sigmoidal layer and a linear output layer with 400 neurons in the hidden layer and 2 neurons in the output layer, computing velocity along x and y axis is considered. The hidden layer can be interpreted as MT cells tuned to different velocities. For training the network, sub sampled features by a factor of 30 from Middlebury sequences are used and the network is trained for 500 epochs using back propagation algorithm till the RMSE of the network over the training samples has reached 0.3. Note that we only have a single network or a regressor and it is applied to all pixels. For training and simulating the experiment PyBrain package has been used.

5.3 Experimental Evaluation and Discussion

Table 5.1 shows the average angular errors (AAE) and the end-point errors (EPE), and the corresponding standard deviations, by considering the Middlebury training set and the Yosemite sequence. Results for the four decoding strategies (intersection of constraints, maximum likelihood, linear decoding with learned weights, and regression using Neural Networks) are reported. Some sample optical flows for the four decoding methods are reported in Figure 5.6. The results show that the intersection of constraints approach gives estimates similar to the ones obtained by considering a linear decoding through learned MT weights. A fitting with Gaussian functions to implement a maximum likelihood decoding does not perform as well as the IOC approach: this is due to the actual MT activity pattern, and to the fact that MT population responses for low speed has several peaks and it is hard to fit a Gaussian.

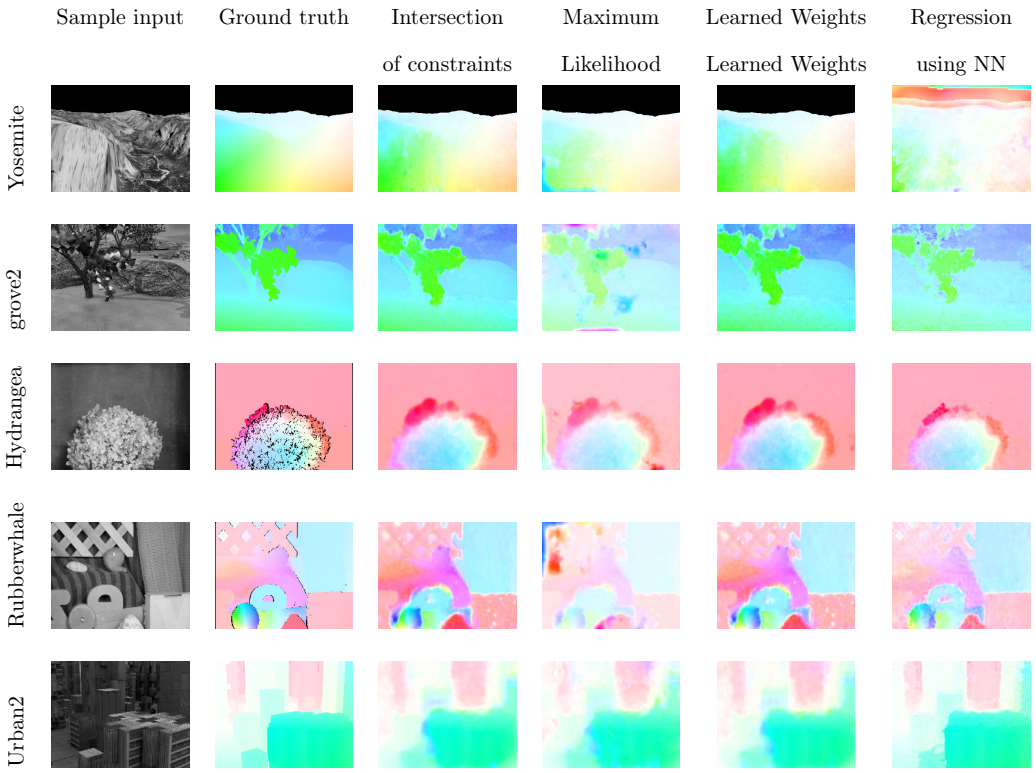


Figure 5.6: Sample results on a subset of Middlebury training set and on the Yosemite sequence.

Observing the results obtained after decoding suggests that scale-space with warping procedure is not well suited for analysis with spatio-temporal features and is inducing larger errors when compared to the regression scheme where the spatio-temporal motion energies across scales are simultaneously taken into consideration.

	Intersection of constraints		Maximum Likelihood		Learned Weights		Regression using NN	
Sequence	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD
grove2	**4.33** ± 10.28	**0.30** ± 0.62	9.78 ± 21.08	0.74 ± 1.30	4.59 ± 9.69	0.32 ± 0.59	5.17 ± 8.49	0.37 ± 0.54
grove3	**9.65** ± 19.02	1.14 ± 1.83	13.73 ± 25.70	1.47 ± 2.32	9.94 ± 18.79	1.15 ± 1.79	9.67 ± 15.39	**1.01** ± 1.42
Hydrangea	5.98 ± 11.19	0.62 ± 0.97	8.88 ± 20.41	0.85 ± 1.44	6.34 ± 11.83	0.65 ± 1.00	**3.22** ± 6.21	**0.29** ± 0.41
RubberWhale	10.16 ± 17.73	0.34 ± 0.54	16.28 ± 26.31	0.73 ± 1.45	10.07 ± 16.65	0.34 ± 0.51	**7.61** ± 8.98	**0.25** ± 0.26
urban2	5.21 ± 10.17	0.58 ± 1.06	14.24 ± 20.37	1.51 ± 1.94	16.46 ± 22.81	1.49 ± 1.91	**4.59** ± 9.69	**0.32** ± 0.59
urban3	15.78 ± 35.94	1.90 ± 3.24	18.24 ± 39.45	1.82 ± 2.91	14.05 ± 33.29	1.74 ± 3.07	**5.76** ± 17.49	**0.80** ± 1.51
Yosemite	**3.49** ± 2.86	**0.16** ± 0.16	5.34 ± 7.24	0.31 ± 0.69	3.80 ± 2.98	0.18 ± 0.18	20.09 ± 14.74	0.86 ± 0.87
all	9.14 ± 16.86	0.85 ± 1.35	12.36 ± 22.94	1.06 ± 1.72	9.32 ± 16.56	0.84 ± 1.29	**8.02** ± 11.57	**0.56** ± 0.80

Table 5.1: Error measurements on Middlebury training set and on the Yosemite sequence.

This is in accordance with earlier model by [Heeger 1988], where plane fitting in spatio-temporal domain has been adapted, indicating that inter-scale interactions are critical in velocity decoding. The neural network based regression has preserved motion edges much better when compared to the warping scheme in most of the sequences, but however it fails in the Yosemite sequence, which indicates that there is some diffusion happening in regions without motion energy as could be seen in the sky region of the optical flow map of the yosemite sequence in Fig. 5.6. The responses of the network need to be more smooth to better match the ground truth, however this is to be expected as this regression scheme does not have any neighbourhood interactions and smoothness criterion in place. This needs to be further investigated by incorporating spatial pooling of the motion energies and spatial interactions at the MT level into the model. On the whole, this indicates that restoring spatial acuity of motion estimation by population decoding is a little studied problem and there is a large scope for improvement as the current decoding schemes do not perform on par with state of the art results in computer vision.

Overall, the results obtained using three different architectures, FFV1MT, APMD and regression based decoding indicate that recurrent interactions both in feature domain and spatial domains play a dominant role in motion integration and ambiguity resolution. Though in APMD model, we incorporate spatial diffusion using trilateral filtering, the feedforward approach typically fails to switch between integration and segregation contextually. This implies that we need to better understand the role played by recurrent interactions about which dynamics could give us a hint.

Part III

Role of recurrent interactions in motion integration

Chapter 6

MT Dynamics

One of the central questions in computational neuroscience is to understand how different forms of neuronal interactions such as feedforward, lateral and feedback interactions play a role in sensory data processing leading to low level percepts. In sensory systems, different computational rules are often evident in different neuronal subpopulations. Considering low level motion, most previous models of motion integration by MT cells [Simoncelli 1998, Rust 2006, Perrone 2008] explain their specific tuning functions by having multiple feedforward inputs, largely ignoring the role of recurrent connectivity, a hallmark of cortical circuits. Therefore they fail to explain the dynamics of these tuning functions and the fact that different behaviours can be achieved by a single subpopulation when varying the spatiotemporal properties of the input [Xiao 2015]. In this chapter, we examine the following questions:

- Can recurrent feature domain interactions between directionally tuned MT cells result in different computational properties such as vector averaging, component selection or component retention behaviour?
- Can the recurrent interactions explain the temporal evolution of the tuning behaviour?
- Can the network attractor properties explain the dynamic transitions in tuning behaviour with respect to the structure of the driving input?

In particular, we study the dependence of the network behaviour with respect to the feature domain center-surround interactions, such centre-surround recurrent mechanisms may be widely applicable to explain contextual modulations in sensory processing.

6.1 Background

Sensory information flows are highly complex and ambiguous with multiple local sensory events occurring simultaneously. Once these events have been accurately extracted, a challenging computational task faced by any sensory system is to integrate or segment them in order to encode behaviourally-relevant information. The difficulty of such a computation is illustrated by visual motion processing. Local motion signals must be selectively integrated in order to reconstruct the direction and speed of a particular surface from the dense and cluttered image flow in oder to overcome the aperture problem [Braddick 1993]. But the same set of signals must also be segregated from the many others that could

belong to other surfaces, multiple surfaces could fall with-in the receptive field of the cell as in the case of motion boundaries or overlapping motion surfaces as in the case of transparency. The rules governing motion integration and segmentation have been extensively investigated at perceptual and physiological levels [Movshon 1985, Qian 1994, Braddick 1997, Treue 2000, Grossberg 2001]. For instance, when presented with two motion directions, or two motion direction distributions, the primate visual motion system can group them according to simple (i.e., vector average) or complex (i.e., intersection-of-constraints) rules. It can also segment them by either suppressing one of the two inputs (i.e., winner-take-all) or by simultaneously representing both of them as in motion transparency. Empirical evidence has been found at both single-cell or population levels for each of these computations in the middle temporal (MT) cortical area of monkeys, a pivotal processing stage in object motion computation [Born 2005]

The classical physiological approach is to identify sub-populations of MT neurons whose behaviours fit with these different computations, thanks to differential weighting (e.g. [Xiao 2015, McDonald 2014, Treue 2000, Movshon 1985]). When presented with plaids made of two superimposed sinusoidal gratings drifting in different directions, some cells encode only one of the two components (component cells) while others encode the pattern motion direction after combining them (pattern cells) [Movshon 1985]. However, recent modelling studies have suggested that these two subpopulations span a continuum along which directionally-selective inputs are differently weighted [Rust 2006]. A similar idea was recently proposed by Xiao and Huang [Xiao 2015] to explain the different types of direction tuning obtained in response to single or bidirectional dot patterns. A majority of MT cells exhibits a single peak tuned to either one of the component or to their mean direction, implementing either a winner-take-all or a vector average computation. Others show two peaks, thus representing the two, overlapping motion directions [Xiao 2015].

There are serveral difficulties with the differential weighting hypothesis. First, one cannot reliably predict the cell responses to bidirectional random dot patterns from their responses to plaids [Xiao 2015]. Second, one must take into account the complex temporal dynamics of the tuning functions that can shift over time from, say the vector average to either the winner-take-all or the transparency solutions [Xiao 2015, Treue 2000]. Indeed, integration and segmentation appear as threads of a complex dynamical computation where inhibition and excitation are shaped adaptively given the spatiotemporal properties of the inputs. Whereas the vast majority of previous theoretical studies on motion integration in area MT have focused on two stage linear-nonlinear computation describing the feedforward scheme [Qian 1994, Rust 2006, Perrone 2008], only a few computational studies have shown that these different cell tuning can be the output of a dynamical system where MT subpopulation are recurrently connected, forming a balanced excitation-inhibition network [Wang 1997, Chance 1999, Ponce-Alvarez 2013]. However, they have not studied in detail the interplay between connectivity and driving input leading to the eventual tuning curves. For instance, [Wang 1997] considers a three layer architecture where V1 cells provide input to MT component cells and MT component cells provide input to MT pattern cells. By consequence, the possibility to make a single sub-population exhibit different kinds of tuning is ruled out. [Ponce-Alvarez 2013] are focussed on noise induced correlations in tuning functions, thus do not study the changes in attractor strength of the solutions induced

due to the structure of the input.

From a computational perspective, understanding the excitatory and inhibitory interactions betweens neurons at the physiological level and groups of neurons at a functional level has been a topic of great interest and has been studied using both discrete [Ellias 1975, Cohen 1983, Majani 1989, Yuille 1989, Cohen 1990, Wolfe 1991, Coultrip 1992, Kaski 1994, Raijmakers 1996, Fukai 1997, Hahnloser 1998, Mao 2007, Arkachar 2007, Martí 2012] and continuous representations [Wilson 1973, Amari 1977, Ben-Yishai 1995, Carandini 1997, Hutt 2003, Coombes 2005] of the abstract feature space.

Under discrete settings, [Ellias 1975] studied on-center off-surround inhibition using shunting type of interactions. They demonstrated various behaviors such as inputing broadening, contrast modulation, peak splitting, formation of spurious peaks etc., under a variety of activation functions with different levels of steepness, excitatory and inhibitory strengths. However, only random/rectangular input activations were considered. One of the popular problem in discrete settings is also the emergence of winner take all (WTA) behavior or k-winner take all (kWTA) behavior, which attempted to identify the number of simultaneously active units at steady state. [Yuille 1989] analytically derived conditions for emergence of WTA behavior for shunting type of interactions. [Cohen 1990] showed the occurrence of Hopf bifurcations in on-center off-surround SNNs with symmetric negative feedback and no self-inhibition. [Ermentrout 1992] has shown that delayed inhibition could lead to oscillatory solutions. [Hahnloser 1998] showed that global inhibition may give rise to multistable WTA mechanism in a recurrent network of neurons but considered a linear activation function above threshold. [Fukai 1997] considered uniform lateral inhibition and showed that ratio of strengths between lateral inhibition and self-inhibition leads to either WTA or winners-share-all behaviors. [Xie 2002] extended the work to a grouping of potential winners in the WTA networks beyond single neuron or uniformly arranged groups of neurons.

In the continuous case, this competition has been studied using neural field equations, these are integro-differential equations describing the time evolution of the population activity operating at a continuum limit. Studies addressed various problems such as formation of bumps and wave patterns explaining hallucinations [Wilson 1973, Amari 1977, Hutt 2003, Coombes 2005], emergence and properties of orientiation selecitivty [Ben-Yishai 1995, Carandini 1997]. However, these models studied the properties of network in terms of their autonomous behaviour with no specific driving input.

These aforementioned studies could be broadly classified into two types, one which derive closed form analytical solutions and others which rely on numerical continuation and bifurcation analysis techniques. Analytical proofs for characterizing and understanding the behaviour of the networks can be quite complicated in the presence of non-linearities, noise and unstructured input. Numerical continuation and bifurcation analysis tools are used to understand the behaviour of these complex networks in situations where closed form analytical solutions can be difficult to derive. Bifurcation analysis identifies the crucial parameters and regimes under which the behaviour of the network changes qualitatively.

[Raijmakers 1996] is one of the few works which reported numerical bifurcation analysis using On-center Off-surround interactions among neurons using shunting neural networks (SNN). However, the input considered there was transient. The

network was primarily used to explore content addressable memory. [Mao 2007] also studied the behavior of recurrent neural networks with lateral inhibition in terms of emergence of Winner Take All behaviours. Bifurcation analysis tools were also considered in continuous settings such as[Ellias 1975]. But both in discrete and continuous settings, neither of the studies did consider structured driving input.

Structured input plays a crucial role as it provides a mapping between stimuli used in experiments to the driving input considered in the models. Such attempts to consider structured driving input in case of understanding motion integration dynamics has been applied in few studies such as [Giese 1998, Rankin 2011, Rankin 2014]. However, these studies are focused at the level perceptual decision making and do not take into account different kinds of stimulus categories and tuning behaviours of different sub-populations.

While these studies depict clear interest in understanding the role of recurrent interactions there are two important aspects that have been ignored so far. There are instances where bifurcation with inputs to neural field models have been considered but no special attention to the shape/characteristics of the input has been paid. For example, studies in memory/winner take all or K-winner take all, usually input is a randomly generated vector. This leaves a wide gap open for understanding specific response characteristics of a network with respect to the characteristics of its specific input. At the same time, it is impractical to characterize the network with respect to all different inputs using bifurcation analysis but it is possible to consider a particular class of stimuli. In this work, we propose to use a linear combination of localized Gaussian bumps as an input class of interest given the ubiquitous evidence for Gaussian like tuning response exhibited by populations of neurons. Apart from the structure of the input another very interesting aspect of the study is the connectivity kernel that governs the strength of local excitation/inhibition. Even though various kinds of connectivities are studied such as uniform lateral inhibition, center-surround inhibition (Mexican Hat) or lateral inhibition with various degrees of self excitation, a parametrized choice of the kernel with associated trade-off has never been established. In this study we attempt to provide insights into the trade-off of choosing a kernel by modelling these three instances as a continuum using a weighted difference of Gaussians.

Herein, using numerical simulations, we study the interplay between connectivity and input to show how it shapes single unit and population tuning.

6.2 Model Description

We model the behaviour of MT neurons using an empirical voltage based ring network that describes the local mean field potential of a group of directionally tuned neurons under different inputs and centre-surround interactions. We primarily focus on the properties of the steady state solutions such as the shape of the direction tuning functions, their number of peaks at convergence and the peak positions with respect to the driving inputs. Let $u(\theta, t)$ denote the activity of the neurons tuned to motion direction $\theta \in [-\pi, \pi)$ and the population dynamics are described by the following neural field equation,

$$\frac{du(\theta,t)}{dt} = -u(\theta,t) + \int_{-\pi}^{\pi} J(\theta-\phi)S(\mu u(\phi,t), th)d\phi + k_i * I_{ext}(\theta),$$

where, J is the connectivity kernel defined as a weighted difference of Gaussians $J_{g_e,\sigma_e,g_i,\sigma_i}(\theta) = g_e G(\theta,\sigma_e) - g_i G(\theta,\sigma_i)$, $G(\theta,\sigma)$: Gaussian function, g_e: excitatory strength, σ_e: extent of excitatory surround, g_i: inhibitory strength, σ_i: extent of inhibitory surround. S is a sigmoid function (μ regulates the sigmoidal gain) and I_{ext} is the driving input representative of the motion stimuli.

At the continuum limit, upon discretization using N samples, the network represents a sub-population of N directionally tuned MT neurons with a smoothly varying directional preference, represented by an angle θ (Fig. 6.1A). Each cell receives afferent input (I_{ext}) from a V1 layer, where stimuli are encoded by Gaussian distributions for each motion input (Fig. 6.1B). The width of the Gaussian bump describes the inherent uncertainity in the local motion estimations at V1 stage. Distributions are broader for gratings than for random dot patterns, reflecting the broader direction-tuning of V1 cells reported in these two conditions [Albright 1984, Mante 2005]. The input is defined by the peak width of each Gaussian distribution (PW) and the peak separation between the two Gaussian distribution (PS), corresponding to a bidirectional motion stimuli. This formulation allows us to encode a variety of stimuli such as gratings, plaids and RDKs with different levels of uncertainities. MT neurons also receive input from the local recurrent interactions (Fig. 6.1C). This local recurrent connectivity depends only on the directional difference, being locally excitatory and laterally inhibitory. It implements a typical centre-surround connectivity kernel in feature space, described by a weighted difference between the Gaussians. Note that we chose to preserve higher order harmonics rather than appealing to a three mode approximation [Curtu 2004, Rankin 2014]. The connectivity kernel is defined by two parameters, the extent of lateral excitation (α) and the strength of the inhibition (β) as illustrated in Figure 6.1C.

6.3 Numerical study of the model

Exploration of the connectivity space with $\tilde{J}_{\alpha,\beta}$

The impact of main parameters governing the nature of local recurrent interactions, the extent of lateral excitation and strength of lateral inhibition are studied using a family of weighted DoG kernels, $\tilde{J}_{\alpha,\beta}(\theta) = J_{g_{e_\alpha},\sigma_{e_\alpha},g_{i_\alpha}+\beta,\sigma_i}(\theta)$. In particular, we study the case of uniform lateral inhibition, so σ_i is fixed to a large value ($\sigma_i = 10\pi$). α is a parameter to smoothly vary the extent of excitatory surround from a narrow (σ_{e_a}) to a broad (σ_{e_b}) bump, depending on a parameter α: $\sigma_{e_\alpha} = \sigma_{e_a} + \alpha\,(\sigma_{e_b} - \sigma_{e_a})$. The other parameters describing the difference of Gaussian functions, g_{e_α} and g_{i_α} are estimated in a closed form with the constraint that, first two Fourier coefficients of $\tilde{J}_{\alpha,0}$, $\widehat{\tilde{J}_{\alpha,0}}[0] = -1$ and $\widehat{\tilde{J}_{\alpha,0}}[1] = 1$, which gives $g_{e_\alpha} = e^{-\frac{\sigma_{e_\alpha}^2}{2}}$ and $g_{i_\alpha} = \frac{1+g_{e_\alpha}}{0.0797}$. β is a free parameter to regulate the strength of lateral inhibition. While considering slow evolution of the inhibition, the inhibition strength g_{i_α} is allowed to evolve slowly governed by the following equation. $g_i(t) = g_{i_{low}} + (g_{i_\alpha} + \beta - g_{i_{low}})(1 - exp(\frac{-t}{\tau_I}))$;

6.3.1 Derivation of the closed form constraints used in $\tilde{J}_{\alpha,\beta}$

Considering weighted Difference of Gaussian kernels $J_{\alpha,\beta}(\theta) = J_{g_{e_\alpha},\sigma_{e_\alpha},g_{i_\alpha}+\beta,\sigma_{i_\alpha}}(\theta)$, where $J_{\alpha,\beta}(\theta) = g_{e_\alpha} w_c(\theta,\sigma_{e_\alpha}) - g_i w_c(\theta,i_\alpha)$ and $w_c(\theta,\sigma)$ is a Gaussian function over

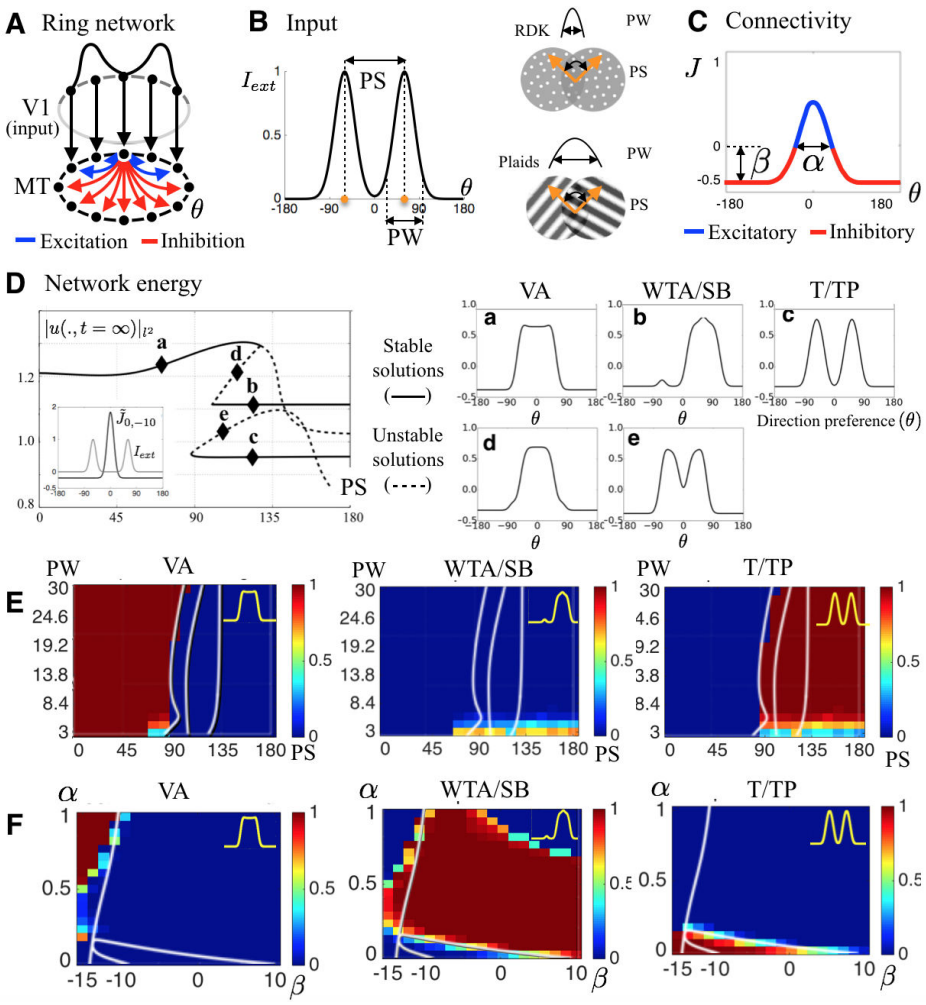


Figure 6.1: The ring model, its different behaviours and the likelihoods of convergence. **(A)** Illustration of ring network modelling visual motion integration at MT cells level, with V1 input and recurrent connections. **(B)** Input defined by two Gaussian distributions parametrized by PS and PW, allowing to represent both RDK and plaids. **(C)** Centre-surround connectivity kernel in feature space parametrized by α and β defining a family of kernels. **(D)** Bifurcation diagram as a function of parameter PS (PW=10, $\alpha = 0$, $\beta = -10$) showing corresponding stable solutions (a,b,c) and unstables ones (d,e) for each branch. Three kinds of solutions co-exist: vector average (VA), winner-take-all (WTA) and transparency (T). Note that referring to [Xiao 2015], winner-take-all will also be designated as side-biased (SB) and transparency as two-peaked (TP). **(E)**, **(F)** show attractor strength of each steady-state solutions (a,b,c), measured as a probability of reaching it from repeated simulations with randomised 100 initial conditions, varying respectively input parameters ($\alpha = 0, \beta = -10$) and connectivity parameters (PS=120, PW=10). Overlaid curves in white are two parameter continuations of the bifurcations indication theoretical transitions in set of stable solutions that co-exist.

the continuous variable θ with a standard deviation σ. Our motif is to derive a set of parameters $(g_{e_c}, \sigma_{e_c}, g_{i_c}, \sigma_{i_c})$ such that when the kernel is discretized between an open internal $[-\pi, \pi)$ with N samples, the first two Fourier coefficients of the

Discrete Fourier Transform $\hat{J}[r]$ of the sampled kernel satisfies the constraints $\hat{J}[0] = 0$ *and* $\hat{J}[1] = 1$ while having a free homotopy variable α which allows us to smoothly transit between kernels having different excitatory widths σ_{e_a} to σ_{e_b}.

We begin by deriving the Fourier coefficients $\hat{J}$. The Fourier transform of a continuous-time Gaussian function of variance σ^2 is also a Gaussian

$$w_c(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\theta^2}{2\sigma^2}} \overset{\mathcal{F}}{\longleftrightarrow} W_c(\Omega) = e^{\frac{-\Omega^2}{2\left(\frac{1}{\sigma^2}\right)}} \tag{6.1}$$

Considering a discrete time Gaussian sequence created by sampling the continuous-time Gaussian function $w_c(\theta)$ at a sampling interval of T

$$w[n] = w_c(nT) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-n^2T^2}{2\sigma^2}} \tag{6.2}$$

The discrete time Fourier transform of $w[n]$ is given by

$$W(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} W_c\left(\frac{\omega - 2\pi k}{T}\right) \tag{6.3}$$

Considering a band limited signal $\frac{1}{\sigma} < \frac{\pi}{3T}$ the DTFT is only significantly contributed by k=0 term, giving the approximation

$$W(\omega) = \frac{1}{T} W_c\left(\frac{\omega}{T}\right) \tag{6.4}$$

Now considering a finite sequence of samples of length M, the Discrete Fourier Transform of the signal is a sample of the DTFT at locations $\omega = \frac{2\pi r}{M}$ for $0 \leq r < M - 1$

$$W[r] = \frac{1}{T} W_c\left(\frac{2\pi r}{MT}\right) \tag{6.5}$$

Considering a weighted difference of Gaussians kernel

$$J_c(\theta) = g_e w_c(\theta, \sigma_e) - g_i w_c(\theta, \sigma_i) \tag{6.6}$$

The FT of this continuous function sampled between the interval $[-\pi, \pi)$ with a sequence of length N then the time period T = $\frac{2\pi}{N}$. Then the DFT of the connectivity kernel is given by

$$\hat{J}[r] = g_e \frac{N}{2\pi} W_c\left(r, \sigma_e\right) - g_i \frac{N}{2\pi} W_c\left(r, \sigma_i\right) \tag{6.7}$$

In case of uniform lateral inhibition We need to define a smooth transition from Kernel J_A defined by the parameters $(g_{ea}, \sigma_{ea}, g_{ia}, \sigma_{ia})$ to kernel J_B defined by parameters $(g_{eb}, \sigma_{eb}, g_{ib}, \sigma_{ib})$. Considering the two dimensional parameter space of excitatory and inhibitory sigmas, then defining a smooth transition amounts to finding the weights (g_{ec}, g_{ic}) that correspond to a chosen point $(\sigma_{ec}, \sigma_{ic})$ on the line joining the two points $(\sigma_{ea}, \sigma_{ia})$ and $(\sigma_{eb}, \sigma_{ib})$ in the parameter space. If we consider the σ_{ia} and σ_{ib} to be large, then by following equation 12 and approximating the Fourier transform of the inhibitory kernel to be a constant we get the Fourier transform of J_A, denoted by $\hat{J}_A$ as

$$\hat{J}_A[r] = \begin{cases} g_{ea} \frac{N}{2\pi} W_c\left(r, \sigma_{ea}\right) - g_{ia} C_1, & \text{if } r = 0 \\ g_{ea} \frac{N}{2\pi} W_c\left(r, \sigma_{ea}\right), & \text{if } r > 0 \end{cases} \tag{6.8}$$

Considering $\sigma_{ia} = \sigma_{ib} >> 2\pi$, then only change in parameter space is governed by excitatory width,

$$\sigma_{ec} = \sigma_{ea} + \alpha(\sigma_{eb} - \sigma_{ea}) \tag{6.9}$$

So now applying the constraints $\hat{J}_C[0] = -1$ and $\hat{J}_C[1] = 1$

$$\hat{J}_C[0] = -1 = g_{ec}\frac{N}{2\pi} - g_{ic}C_1 \tag{6.10}$$

$$\hat{J}_C[1] = 1 = g_{ec}\frac{N}{2\pi}W_c\left(r = 1, \sigma_{ec}\right) \tag{6.11}$$

It implies

$$g_{ec} = \frac{1}{\frac{N}{2\pi}W_c\left(r = 1, \sigma_{ec}\right)} \tag{6.12}$$

and

$$g_{ic} = \frac{1 + g_{ec}\frac{N}{2\pi}}{C_1} \tag{6.13}$$

Parameter values, initial conditions and numerical computations

The parameters used for the numerical simulations are gathered in Table 6.1. In case of simulations without noise, standard ode solver, ODE23T is used with absolute tolerance value set to 10^{-12}. For the simulations with noise, Euler Maruyama method is being used. In the simulations initial conditions are set to a low level of random activity. In order to carry out numerical continuation and bifurcation analysis, Auto07p package is used, allowing us to track the bifurcation point in one and two-dimensional parameter space. For bifurcation analysis computations are carried out in the absence of noise for a variety of combination of driving input and connectivity kernels. The feature space is discretized into 404 samples and values of the coefficients used are chosen so that the equation is at continuum limit.

Results

6.3.2 Network behaviour

Using numerical bifurcation analysis we first identified the stable solutions and the parameter regimes over which these different solutions coexist. When the network is stimulated with a bidirectional input, varying the distance between the two input peaks (PS) leads to different types of solution. These are shown in a bifurcation diagram in Fig. 6.1D, where stable solution branches are solid and unstable branches dashed. The tuning curves of the three stable solutions are indicated in the right panels and corresponded to the three main cell tuning already reported[Xiao 2015]. Cases a, b and c correspond to the observed vector average (VA), winner-take-all (WTA) and transparency (T) tuning functions, respectively. For small PS value, VA is the only possible solution. For large PS value, WTA or T coexist. For an intermediate range ($95 \lesssim PS \lesssim 130$), we identify a critical operating regime for the model, where it is capable of producing the three prototypical tuning types in one small parameter region. We observed that the WTA solution was side-biased, indicating that the other, competing input was not fully suppressed, similar to the empirical evidence in macaque area MT [Xiao 2015]. Unstable solution branches

Description	Parameter	Value
Number of samples in $[-\pi, \pi)$	N	404
Sigmoid threshold	th	3.0
Sigmoid gain	λ	16.0
Input gain	k_i	0.1
Population time constant	τ_p	1.0, 5.0, 10.0
Inhibition time constant	τ_I	30-100
Homotopy variable to regulate excitation width	α	[0,1]
Inhibition offset	β	[-10,15]
Excitation width	$\sigma_{e_\alpha} = \sigma_{e_a} + \alpha\,(\sigma_{e_b} - \sigma_{e_a})$	[11.5°(σ_{e_a}), 60°(σ_{e_b}))
Inhibitory width	σ_i	1800(>> 360)
Excitation strength	g_{e_α}	$e^{-\frac{\sigma^2_{e_\alpha}}{2}}$
Inhibition strength	g_{i_α}	$\frac{1+g_{e_\alpha}}{0.0797}$
Peak Separation	PS	(0, 180]
Peak width	PW	5°-30°

Table 6.1: Parameter values used in the numerical studies

(d,e) can link the stable branches, but are not critical in this study. Stable solution branches terminate at bifurcations. The intersection of a and d is a pitchfork bifurcation, of d and b a fold bifurcation and of c and e a fold bifurcation. These can be tracked in terms of two parameters (Fig. 6.1E) to demarcate entire parameter regions where different solution types exist (a two-parameter phase diagram for the model).

The network behaviour is best characterized by maps of attractor strength that reveal which of the stable solutions identified in the bifurcation analysis dominates. Attractor strength of each solution, measured as a probability of reaching it from repeated simulations with randomised initial conditions, was computed at every combination external parameters (input: PS,PW) and internal parameters (connectivity kernel: α, β) (Fig. 6.1E,F). Although in a given parameter regime two or more types of solutions can coexist as stable solutions, the network might be much more likely to converge to one of these from a random initialisation. In, Fig 6.1E, for large peak separation, the transparency case dominated over a large range of peak width. Conversely, narrow peaks that are widely separated yielded to the WTA solution. The likelihood of the VA solution was maximal for small peaks separation, regardless of the PW. In Fig 6.1F, we observe that maintaining two peaks at population level requires both low inhibition and narrow excitation whereas VA requires low inhibition but broad excitation. The WTA solution is highly probable across a wide range of inhibition strength and excitation extent. Thus, the emergence of the different neural solutions to the motion integration and segmentation problem results from the interplay between the properties of the inputs and the shape of the centre-surround interactions. Moreover, these different types of direction tuning are solutions of a single dynamical system.

6.3.3 Local recurrent interactions lead to prototypical tuning behaviours found in MT

Recording single unit activity in macaque area MT, Xiao and Huang [Xiao 2015] investigated the neuronal responses to random dot patterns made of either one or two motion direction inputs. Interestingly, they documented prototypical behaviours as shown in Fig. 6.2A. For a fixed direction difference between motion directions of a dot pattern, the authors found three different classes of cells. Type A1 represents the vector average (VA) of the two motion inputs while A4 superposes them, yielding to a two peaked tuning function (TP). The two other examples (A2, A3) form a single class that suppress one or the other of the two inputs. These side-biased (SB) cells implement a weak winner-take-all computation where an influence from the suppressed inputs can still be seen. Here we show that a population of direction selective cells can produce these prototypical behaviours under the influence of local recurrent interactions.

We simulated the network by sequentially varying the pattern direction of the driving input. Figure 6.2B, illustrates the MT population direction preference, as a function of the average input motion direction. Four cases are presented, corresponding to different parameters of the input and different recurrent connectivity regime (Fig. 6.2C). Sampling the population vertically or horizontally provides the MT population tuning (Fig. 6.2D) and a single cell tuning (Fig. 6.2E). Column B1 shows that the network accurately represents a unidirectional input. Column B4 illustrates the population and single cell tuning for two widely separated inputs (120deg), for the same recurrent connectivity, with a low inhibition regime ($\beta = -10$). The two peaks are preserved, yielding to a bimodal tuning function. Notice that each peak is now sharper than observed with a single direction input. Narrowing the input direction difference to 90deg (clolumn B2) changed the tuning functions at both population and single-cell levels, with now the vector average being represented. Thus, as reported by [Xiao 2015] and others [McDonald 2014], VA and TP solutions can be achieved, depending on the spatiotemporal properties of the input (Fig. 6.2, compare B4 with A4 and B2 with A1). Using the same input as in B4, column B3 illustrates that the network shifted to a winner-take-all (WTA) solution when inhibition strength is increased ($\beta = 10$). Here an input bias to the $-60°$ direction was incorporated (Fig. 6.2D, column B3, orange curve). In our model, a high inhibition regime with bias introduced through afferent connection strengths leads to a robust side-biased behaviour. A similar case with symmetric input is considered in Fig. 6.3B. Notice however that a small response to the suppressed direction is still evident, due to the local recurrent excitation. Single cell activities are more variable than in the population activity. This can happen when the network is operating in a parameter regime where different solutions co-exist with one attractor largely (but not totally) dominating and the noise in the input driving the network through the possible stable solutions(see Figure 6.1).

6.3.4 Attractor bias leads to consistent tuning and predicts fluctuations in component selection

One of the hallmarks of recurrent interactions is the existence of multiple stable solutions for the same set of parameters albeit with different attraction strengths.

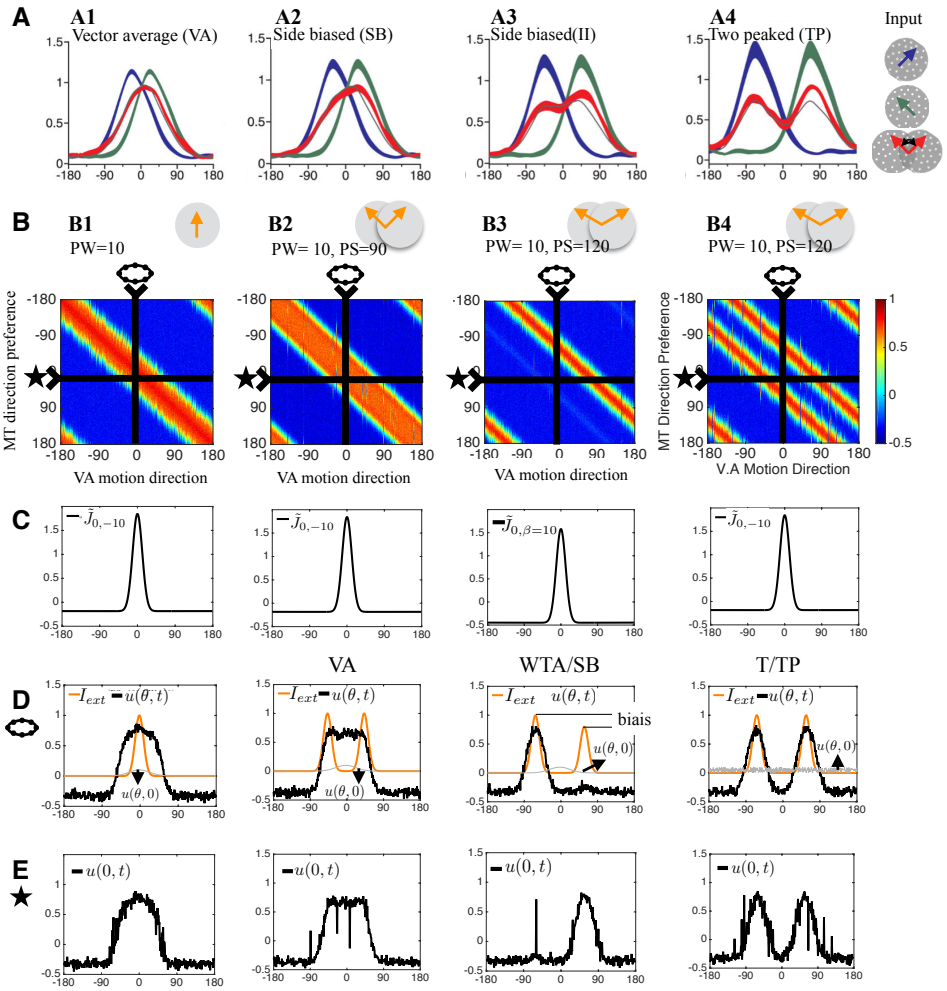


Figure 6.2: Local recurrent interactions lead to prototypical tuning behaviours found in MT. **(A)** Four examples of prototypical single cell tuning behaviours found in MT (Adapted from Xiao and Huang, J. of Neuroscience 2015 [Xiao 2015]. Blue and green curves show the tuning functions obtained when presented with a dot pattern moving coherently in one of the two directions. Red curves show the tuning functions obtained with the two direction components overlapped, forming a bidirectional pattern. **(B)** Simulated the network by sequentially varying the pattern direction of the driving input for a uni-directional stimulus (column B1) and bi-directional ones (columns B2–B4). **(C)** Connectivity kernels used in each cases. **(D)** MT population tuning obtained from B by doing a vertical cut. **(E)** Single MT cell tuning obtained from B by doing an horizontal cut.

Under this multi-stability, the network could converge to any of the stable solutions states and it would manifest as a random fluctuations at a single cell level. Observations pertaining to stable tuning behaviour could be emerging from the regimes where one of the attractor is dominant either due to parameter regime or structure of the driving input. For example, strong affinity towards vector average or two-peaked solution could be due to the network operating at regime where VA or two peaked solutions are the dominant attractors. The steady tuning of side bias could

be due to a slight asymmetry in the afferent and network operating in a winner take all domain, the component of the input with lower amplitude would be suppressed. In feedforward models, this kind of component suppression would require a huge amount of asymmetry in the feedforward inputs. Considering recurrent interactions, even a slight bias would result in consistent suppression of one of the components.

One prediction of our network model is that tuning functions could exhibit fluctuations as the network could select either one of the components when stimulated with input having purely symmetric component strength. This is evident in the bifurcation diagram shown in Figure 1B, where different computational solutions can coexist for a given input range. Therefore, we presented repeated trails with bidirectional inputs having different motion direction difference and/or various degrees of strength difference between the two motion components. In experimental studies, this latter manipulation would correspond to selectively decreasing the signal-to-noise ratio for one component. All other parameters regarding recurrent connectivity were kept constant. Fig. 6.3.A shows that population direction tuning remains largely unchanged when the angular difference between the two component motions was small ($-60°$) and the relative strength was varied. In Fig.. 6.3.B however, the population tuning was remarkably stable with asymmetric inputs but was highly fluctuating when the two motion components have the same strength.

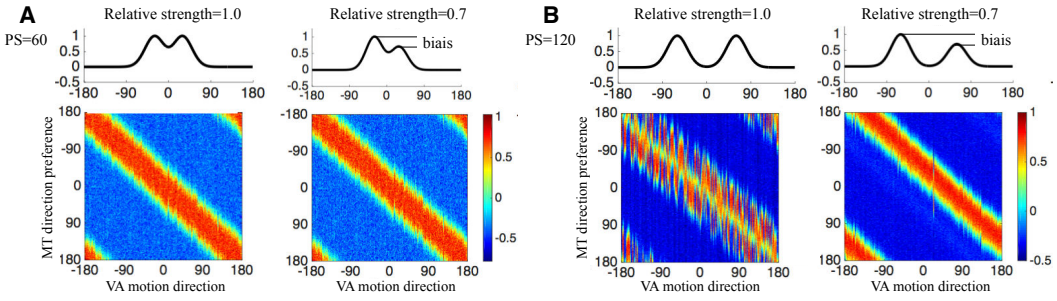


Figure 6.3: Tuning behaviour changes with respect to the separation of the components and asymmetry in the input. We simulated the network by sequentially varying the pattern direction of the driving input as in Fig. 6.2B, with **(A)** PS=60, and **(B)** PS=120. For both cases, a comparison is made between symmetric and asymmetric input. Other parameters are fixed (PW=20, $\alpha = 0$, $\beta = -5$).

6.3.5 Predicting tuning behaviour for different spatiotemporal properties of bidirectional motion inputs

One remarkable properties of pattern motion integration in macaque area MT is that cell responses to bidirectional plaid patterns cannot be fully predicted from bidirectional random dot patterns, and reciprocally [Xiao 2015, McDonald 2014]. The transitions in the behaviour of the cells classified to be side biased, two peaked or averaging using RDKs when stimulated by plaids are presented in Fig. 6.4.D. Recurrent interactions may play a role in shaping the cell response properties to the exact spatio-temporal properties of the motion input.

In our model, we can examine these translations by examining the attractor strength of the stable solutions across variety of internal parameters allowing us

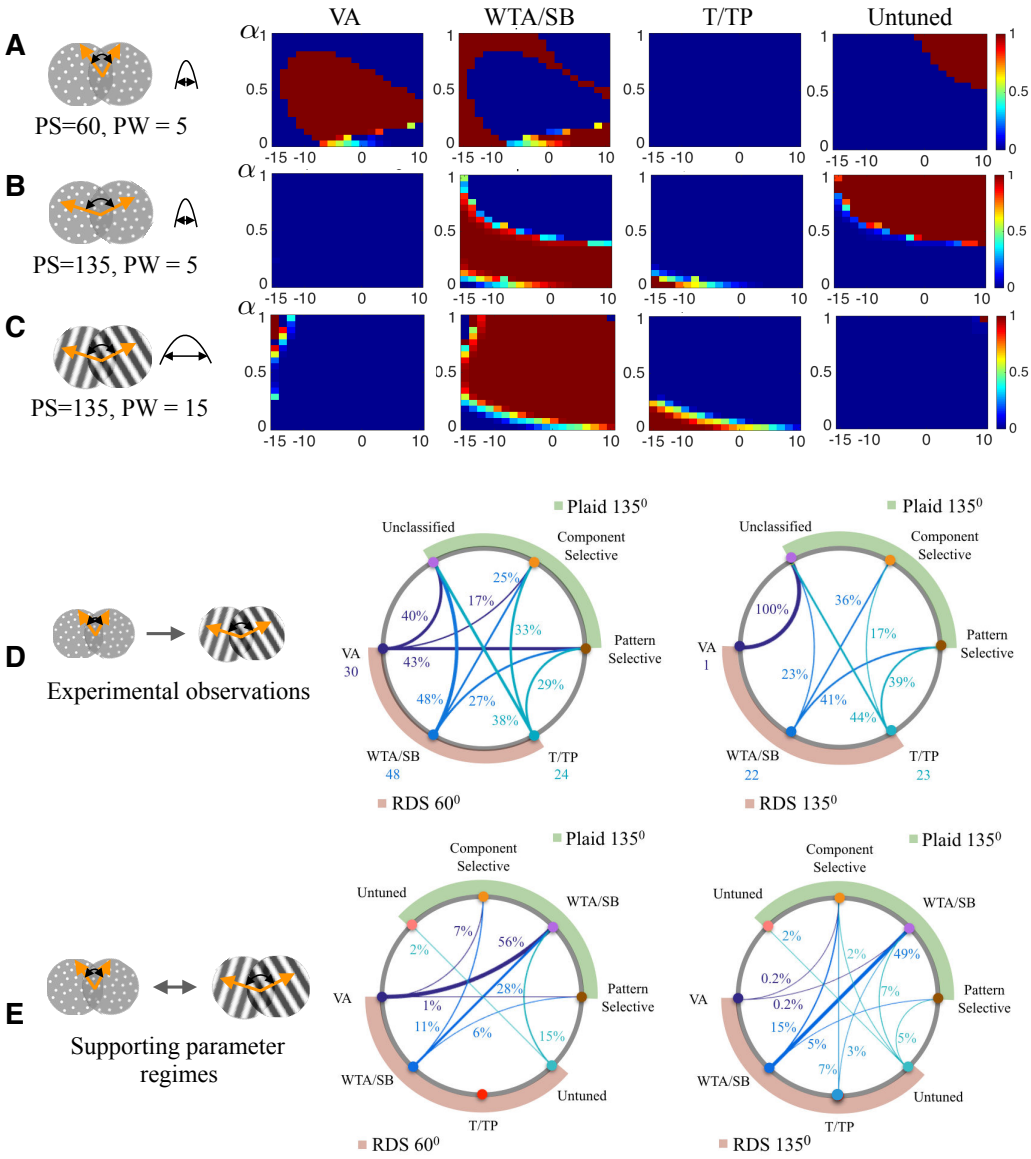


Figure 6.4: Predicting tuning behaviour with respect to stimuli. A,B,C: Convergence probabilities to stable solutions when the network is driven with inputs representative of plaid and RDK stimuli. The maps indicate the probability of convergence to particular solution when the network is simulated with the input using connectivity parameters (α, β), measured using 100 trials. D: Transitions in the tuning behaviour of the cells that exhibited a particular type of tuning behaviour when tested using RDKS to Plaid stimuli (Data from Table. 2 of [Xiao 2015]). E: The percentage of overlap in the parameter regimes that support specific tuning behaviour when the network is driven with input representative of the two stimuli. This overlap is indicative of the plausible transitions in the tuning behaviour of the cell with respect to input.

to visualize how the attractor strengths would shift from one stimulus category to other by simple superposition as illustrated in Fig. 6.4. We started by measuring the attractor strength of each solution at different connectivity regimes from repeated

trials (100) using the specific input. We then constructed a likelihood map from this condition. Once the likelihood maps for all the solution types across different input conditions are obtained, we identified the connectivity regimes which could support transitions in one type of tuning for a particular input to other types of tuning with a different input by overlaying the thresholded maps. The percentage of the overlap within the range of parameters tested is presented in Fig. 6.4.E. Interestingly, some of the experimentally observed transitions could not be explained by changes in the driving input alone: for example transition from transparency to vector averaging cannot be observed without a shift in the overall excitatory and inhibitory structure towards a broader excitatory extent and stronger inhibition. This could potentially indicate a systemic contextual modulation in motion integration driven by form based cues. Another broad trend that has been observed is that transparency, or co-existence of two peaks is supported by local ambiguity. In case of sharp input, the likelihood of two peaks is very little. At the same time, the likelihood of VA increases with increased ambiguity. This could indicate that, when ambiguity is low spatially a salt and pepper kind of tuned populations selecting a particular component could be used the by system to represent transparency. When such stimuli are spatially proximal, it could lead to fluctuating perception.

6.3.6 Temporal dynamics

Experimentally it has been shown that neurons develop the selectivity for a particular direction after the 60-80ms of time delay [Smith 2005, Xiao 2015]. Here we tested the hypothesis that slow onset of inhibition could lead to such temporal evolution of the selectivity and transition in the tuning behaviour. To do so, the inhibition strength (β) is allowed to evolve temporally on a slower time scale. The intuition behind this observation is that uninhibited lateral interactions could lead to integration early on developing a broadly tuned response and later on the activity can be shaped by the inhibition. The results obtained using two different input configurations in confirmation with this hypothesis can be found in Fig. 6.3.6. As expected, early lateral interactions allowed the spread of the activity leading to broad tuning. The onset of inhibition later on shaped the tuning to resemble one of the prototypical solutions based on the the eventual strength it converges into. This could hint at a possibility of early integration of sensory inputs not only in vision but also in other modalities.

More interestingly, our model could not find a transition from a broad VA like response to two peaked response that was reported by [Xiao 2015]. The transitions reported could be seen in Fig. 6.6. In case of RDK stimuli whose motion directions are separated by 60^0, the temporal evolution of two peaked behaviour from a broad untuned response has been reported Fig. 6.6.b, however, our model could not find this type of transition. There are two potential causes for it. Firstly, the likelihood of two peaked solution is very low when the driving input considered was sharp as in the case of RDKs. So, the model with highly likelihood jumps from a Vector Average solution branch to side bias (WTA) branch. Even if we set inhibition such that only vector average or two peaked solutions are only stable solutions, it takes a lot of noise for the system to destabilize from VA solution and reach a two-peaked solution branch. Second, there is qualitative change in the early tuning that has been reported when two motion components considered are closer (60^0)

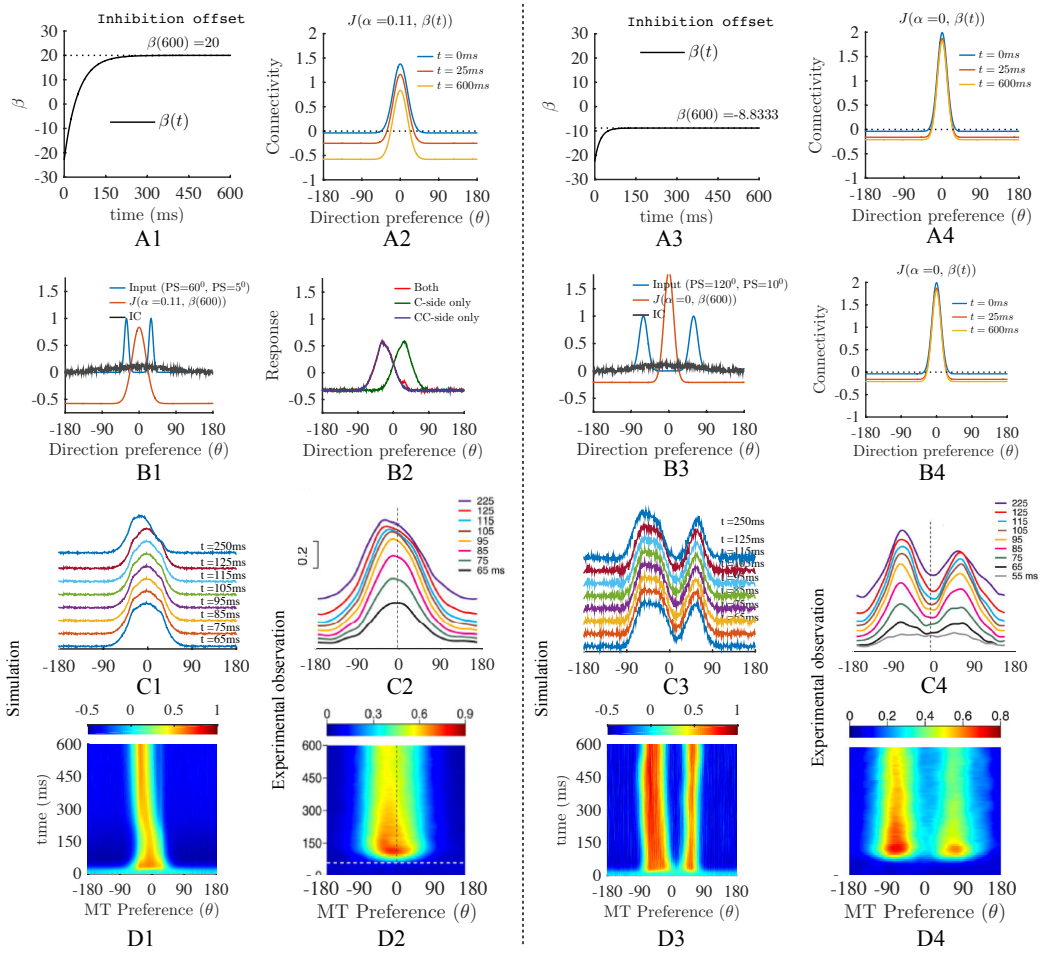


Figure 6.5: Temporal tuning behaviour. A-F (1,2) illustrate the system exhibiting side bias. A-F (3,4) illustrate the system exhibiting two peaked behaviour. A1, A3: evolution of lateral inhibition strength. A2, A4: connectivity kernels sampled at different time intervals. B1, B3: Input and initial conditions of the network. B2, B4: Population tuning at convergence. C1, C3: Early tuning behaviour observed in simulations. C2, C4: Early tuning behaviour observed in experimental recordings. D1, D3: Temporal dynamics observed in simulations. D2, D4: Temporal dynamics experimentally observed with RDKs [Xiao 2015].

versus farther apart in direction space $(90^0, 135^0)$. This can be readily observed by comparing Fig. 6.6.a and 6.6.d. In 6.6.a, the activity seems to spread from the center or VA direction and in case of 6.6.d, the activity seems to spread from the constituent component directions. The reason for this qualitative shift is unknown and needs further investigation.

6.4 Discussion

We have shown that a variety of tuning behaviours observed in the sub-populations of macaque direction selective MT neurons could be explained by a recur-

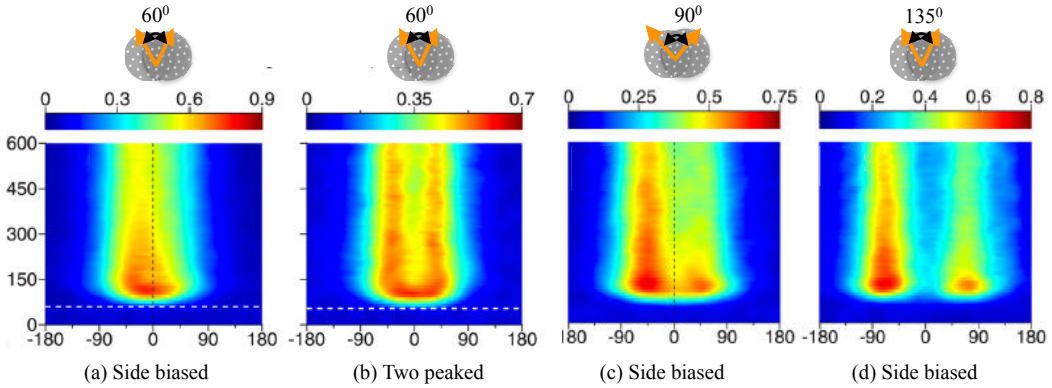


Figure 6.6: Illustrating temporal evolution of tuning. a,c,d: show side biased tuning behaviour evolving from an early broad response. b: shows two peaked behaviour evolving after an early untuned reponse, our model does not capture this transition. Figure adapted from [Xiao 2015].

rently interacting group of cells tuned to different directions. Previous studies have attributed the different tuning behaviours to functionally different subpopulations [Movshon 1985, Rust 2006, McDonald 2014]. Even though we do not refute the existence of distinct cell types, using local feature domain recurrent interactions we have not only reproduced different types of tuning but can explain change of tuning depending on the input type. Experimental evidence suggests that tuning behaviour emerges temporally, [Smith 2005, Pack 2001] have demonstrated that pattern selectivity emerges over a delay of 60-70 ms, [Xiao 2015] have shown that component selectivity could emerge after an initial preference for vector averaging. These observations have been modelled earlier using functional feedforward models such as [Rust 2006, Perrone 2008] but these models require considering cells with different temporal sensitivity and also cannot necessarily link the computational elements such as tuned versus untuned normalization or strong feedforward inhibition that model parameters demand. Local recurrent have so far received little attention despite being speculated by [Smith 2005, Xiao 2015] in this chapter we have demonstrated that local recurrent interactions can explain the temporal emergence of tuning behaviours.

Canonical competition at perceptual and neuronal levels/Theoretical link between cortical responses and perception

Considering the feature space local recurrent interactions could be a step in establishing the connection between low-level neural dynamics associated with sensory processing and perceptual dynamics, including transparency. This requires us to understand the behaviour to be expected under different kinds of interactions. Our study, establishes that competitive recurrent interactions in the feature space that have been explored earlier at macroscopic(perceptual level) [Rankin 2014] earlier can account for different tuning behaviours observed at the neuronal level [Xiao 2015]. These local recurrent interactions could be a key to connect low level neuronal dy-

namics with high level perceptual dynamics.

Contextual modulation leads to multiple behaviours from the same network

It is well known in the experimental literature that directional tuning can vary significantly based on the type of the stimuli used, for example Type 1 plaid, Type 2 plaid or Random Dot Kinematograms. The pattern/component cells respond differently to RDKs, which has been demonstrated by [Xiao 2015]. These kinds of dynamic changes in the behaviour can be attributed to both changes in the spatio-temporal input representative of local ambiguity as well as the contextual modulations that can occur in the network in terms of lateral excitation and inhibition strength. Our model provides a systemic way to investigate role of each of these factors. Our simulations facilitate prediction of most likely changes in the tuning under same context and also lets us estimate contextual changes (internal parameters) the network has to undergo such as improved extent of lateral excitation or increased inhibition strength to explain the observations.

Multi-stability and unclassified response

Recurrence as observed could lead to multi-stability. Multi-stability at sub-population levels could manifest itself as fluctuations at single cell across trials, leading to a lack of specific tuning behaviour or high variance in cells response across trials. This is a potential factor for explaining the cause for unclassified cells seen in the MT area. There could be a potential sub-group which fluctuates on component selection under symmetric driving input, which could be stabilised by changing the contrast in one of the driving components. This is a prediction that could be tested experimentally. The effective internal parameters could be changing with respect to stimulus type, for example RDKs versus Plaids.

General theoretical results for a ring network

We have presented a detailed numerical study that elucidated the behaviour of a ring network with neurons interaction in a on-center off-surround manner, going beyond traditional characterization of the network in terms of the solution space in autonomous mode we have proposed a structured driving input which facilitated the representation of stimuli used in physiology and psychophysics. Our proposal reduced the study into four critical variables, two of them reflective of external attributes that could be controlled, peak width and peak separation and two internal parameters, excitatory extent and inhibition strength. The bifurcation analysis gives us an idea of the stable solutions and parameter regimes where each of the solutions are stable, a complementary stochastic trials indicate the strength of the stable solutions. This computational characterization of the network is broadly applicable for modelling studies and can be used to model other experimental observation in identical scenarios.

6.5 Conclusion

Recurrent local interactions in the feature domain can reproduce variety of tuning behaviours that have been reported in literature. Interestingly, proto-typical tuning curves that have been reported form the stable solutions of the ring network under center-surround connectivity. The attraction strength of different stable solutions can potentially explain the dynamic changes in the tuning behaviour that has been observed physiologically. These recurrent interactions could potentially form a bridge between the models of motion integration and models that capture transparency as they eliminate the need for different sub-populations and strong competition. These recurrent interactions need to be further investigated using spatialized version for developing generic models which can work under different stimulus categories.

Part IV

Towards synergistic models in vision...

Chapter 7

Task centric exploration of biological and computer vision

The basic premise of the thesis is that scaling up models in biological vision would be mutually beneficial for both computer and biological vision. In parts II and III, we have shown that evaluating models rooted in biological vision taking a task centric view gave us insights to better constrain the models and explore the role of recurrent interactions. A natural question to ask is how could computer vision benefit from studies in biological vision. The goal of this chapter is to examine how novel computer vision approaches could be developed from the biological insights. It is a manifesto for developing and scaling up models rooted in experimental biology (neurophysiology, psychophysics, etc.) leading to an exciting synergy between studies in computer vision and biological vision. Our conviction is that the exploding knowledge about biological vision, the new simulation technologies and the identification of some ill-posed problems have reached a critical point that will nurture a new departure for a fruitful interdisciplinary endeavour. The resurgence of interest in biological vision as a rich source for designing principles for computer vision is evidenced by recent books [Petrou 2008, Frisby 2010, Hérault 2010, Pomplun 2012, Cristobal 2015, Liu 2015] and survey papers [Tsotsos 2014, Cox 2014]. However, we feel that these studies were more focused on computational neuroscience rather than computer vision and, second remain largely influenced by the hierarchical feedforward approach, thus ignoring the rich dynamics of feedback and lateral interactions.

This chapter is organised as follows. In Sec. 7.1, we revisit the classical view of the brain as a hierarchical feedforward system [Kruger 2013]. We point out its limitations and portray a modern perspective of the organisation of the primate visual system and its multiple spatial and temporal anatomical and functional scales. In Sec. 7.2, we appraise the different computational and theoretical frameworks used to study biological vision and re-emphasise the importance of putting the task solving approach as the main motivation to look into biology. In order to relate studies in biological vision to computer vision, we focus in Sec. 7.3 on three archetypal tasks: sensing, segmentation and motion estimation. These three tasks are illustrative because they have similar basic-level representations in biological and artificial vision. However, the role of the intricate, recurrent neuronal architecture in figuring out neural solutions must be re-evaluated in the light of recent empirical advances. For each task, we will start by highlighting some of these recently-identified biological mechanisms that can inspire computer vision. We will give a structural view of these mechanisms, relate these structural principles to prototypical models from both biological and computer vision and, finally we will detail potential insights and perspectives for rooting new approaches on the strength of both fields. Finally, based on the prototypical tasks reviewed, we will propose in Sec. 7.4, three ways to identify which studies from biological vision could be leveraged to advance computer

vision algorithms.

7.1 Deep cortical hierarchies?

7.1.1 The classical view of biological vision

The classical view of biological visual processing that has been conveyed to the computer vision community from visual neurosciences is that of an ensemble of deep cortical hierarchies (see [Kruger 2013] for a recent example). Interestingly, this computational idea was proposed in computer vision by David Marr [Marr 1982] even before its anatomical hierarchy was fully detailed in different species. Nowadays, there is a general agreement about this hierarchical organisation and its division into parallel streams in human and non-human primates, as supported by a large body of anatomical and physiological evidences (see [Ungerleider 1994, Van Essen 2003, Markov 2013] for reviews). Fig. 7.1(a)–(b) illustrates this classical view where information flows from the retina to the primary visual cortex (area V1) through two parallel retino-geniculo-cortical pathways. The magnocellular (M) pathway conveys coarse, luminance-based spatial inputs with a strong temporal sensitivity towards Layer 4Cα of area V1 where a characteristic population of cells, called stellate neurons, immediately transmit the information to higher cortical areas involved in motion and space processing. A slower, parvocellular (P) pathway conveys retino-thalamo-cortical inputs with high spatial resolution but low temporal sensitivity, entering area V1 through the layer 4Cβ. Such color-sensitive input flows more slowly within the different layers of V1 and then to cortical area V2 and a network of cortical areas involved in form processing. The existence of these two parallel retino-thalamo-cortical pathways resonated with neuropsychological studies investigating the effects of parietal and temporal cortex lesions [Ungerleider 1982], leading to the popular, but highly schematic, two visual systems theory [Ungerleider 1982, Ungerleider 1994, Milner 2008] in which a dorsal stream is specialised in motion perception and the analysis of the spatial structure of the visual scene whereas a ventral stream is dedicated to form perception, including object and face recognition.

At the computational level, the deep hierarchies concept was reinforced by the linear systems approach used to model low-level visual processing. As illustrated in Fig. 7.1(c), neurons in the primary visual system have small receptive fields, paving a high resolution retinotopic map. The spatiotemporal structure of each receptive field corresponds to a processing unit that locally filters a given property of the image. In V1, low-level features such as orientation, direction, color or disparity are encoded in different sub-populations forming a sparse and overcomplete representation of local feature dimensions. These representations feed several, parallel cascades of converging influences so that, as one moves along the hierarchy, receptive fields become larger and larger and encode for features of increasing complexities and conjunctions thereof (see [DeYoe 1988, Roelfsema 2000] for reviews). For instance, along the motion pathway, V1 neurons are weakly direction-selective but converge onto the medio-temporal (MT) area where cells can precisely encode direction and speed in a form-independent manner. These cells project to neurons in the median superior temporal (MST) area where receptive fields cover a much larger portion of the visual field and encode basic optic flow patterns such as rotation, translation

or expansion. More complex flow fields can be decoded by parietal neurons when integrating these informations and be integrated with extra-retinal signals about eye movements or self-motion [Bradley 2008, Orban 2008]. The same logic flows along the form pathway, where V1 neurons encode the orientation of local edges. Through a cascade of convergence, units with receptive fields sensitive to more and more complex geometrical features are generated so that neurons in the infero-temporal (IT) area are able to encode objects or face in a viewpoint invariant manner (see Fig. 7.1(c)).

Object recognition is a prototypical example where the canonical view of hierarchical feedforward processing nearly perfectly integrates anatomical, physiological and computational knowledges. This synergy has resulted in realistic, computational models of receptive fields where converging outputs from linear filters are nonlinearly combined from one step to the subsequent one [Cadieu 2007, Nandy 2013]. It has also inspired feedforward models working at task levels for object categorisation [Serre 2007b, Serre 2007a] as illustrated in Fig. 7.1(d), prominent machine learning solutions for object recognition follow the same feedforward, hierarchical architecture where linear and nonlinear stages are cascaded between multiple layers representing more and more complex features [Hinton 2006b, Cox 2014].

7.1.2 Going beyond the hierarchical feedforward view

Despite its success in explaining some basic aspects of human perception such as object recognition, the hierarchical feedforward theory remains highly schematic. Many aspects of biological visual processing, from anatomy to behaviour, do not fit in this framing. Important aspects of human perception such as detail preservation, multi-stability, active vision and space perception for example cannot be adequately explained by a hierarchical cascade of expert cells. Furthermore, taking into account high-level cognitive skills such as top-down attention, visual cognition or concepts representation needs to reconsider this deep hierarchies. In particular, the dynamics of neural processing is much more complex than the hierarchical feedforward abstraction and very important connectivity patterns such as lateral and recurrent interactions must be taken into account to overcome several pitfalls in understanding and modelling biological vision. In this section, we highlight some of these key features that should greatly influence computational models of visual processing. We also believe that identifying some of these problems could help in reunifying natural and artificial vision and addressing more challenging questions as needed for building adaptive and versatile artificial systems which are deeply bio-inspired.

Visual processing starts at the retina and the lateral geniculate nucleus (LGN) levels. Although this may sound obvious, the role played by these two structures seems largely underestimated. Indeed, most current models take images as inputs rather than their retina-LGN transforms. Thus, by ignoring what is being processed at these levels, one could easily miss some key properties to understand what makes the efficiency of biological visual systems. At the retina level, the incoming light is transformed into electrical signals. This transformation was originally described by using the linear systems approach to model the spatio-temporal filtering of retinal images [Enroth-Cugell 1984]. More recent research has changed this view and several cortex-like computations have been identified in the retina of

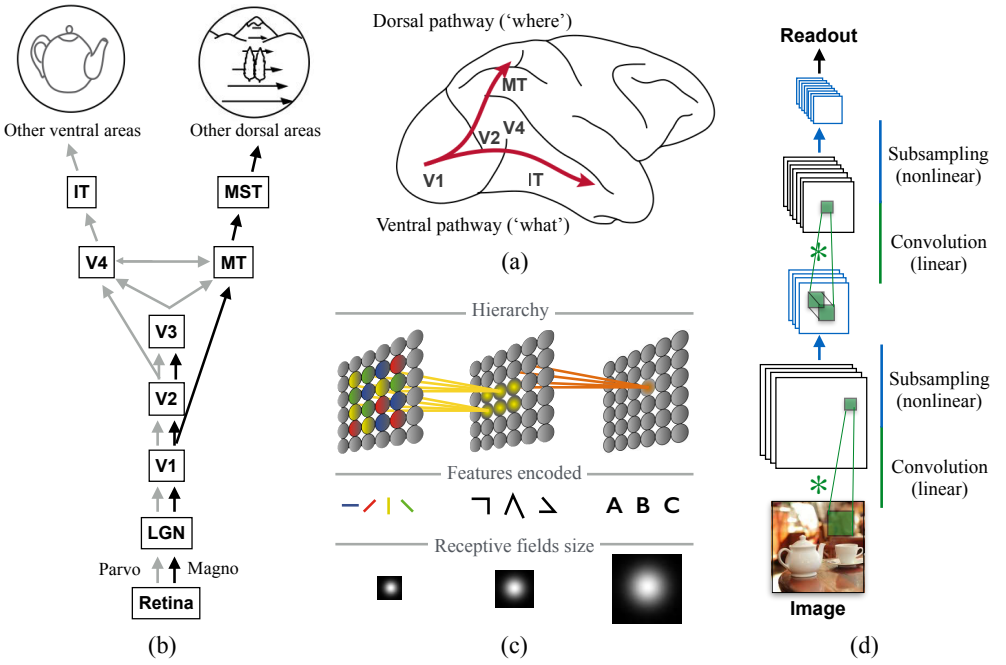


Figure 7.1: The classical view of hierarchical feedforward processing. (a) The two visual pathways theory states that primate visual cortex can be split between dorsal and ventral streams originating from the primary visual cortex (V1). The dorsal pathway runs towards the parietal cortex, through motion areas MT and MST. The ventral pathway propagates through area V4 all along the temporal cortex, reaching area IT. (b) These ventral and dorsal pathways are fed by parallel retino-thalamo-cortical inputs to V1, known as the Magno (M) and Parvocellular pathways (P). (c) The hierarchy consists in a cascade of neurons encoding more and more complex features through convergent information. By consequence, their receptive field integrate visual information over larger and larger receptive fields. (d) Illustration of a machine learning algorithm for, e.g., object recognition, following the same hierarchical processing where a simple feedforward convolutional network implements two bracketed pairs of convolution operator followed by a pooling layer (adapted from [Cox 2014]).

different vertebrates (see [Gollisch 2010, Kastner 2014] for reviews, and more details in Sec. 7.3.1). The fact that retinal and cortical levels share similar computational principles, albeit working at different spatial and temporal scales is an important point to consider when designing models of biological vision. Such a change in perspective would have important consequences. For example, rather than considering how cortical circuits achieve high temporal precision of visual processing, one should ask how densely interconnected cortical networks can maintain the high temporal precision of the retinal encoding of static and moving natural images [Field 2007], or how miniature eye movements shapes its spatiotemporal structure [Rucci 2015].

Similarly, the LGN and other visual thalamic nuclei (e.g., pulvinar) should no longer be considered as pure relays on the route from retina to cortex. For instance, cat pulvinar neurons exhibit some properties classically attributed to cortical cells, as such pattern motion selectivity [Merabet 1998]. Strong centre-surround interactions have been shown in monkeys LGN neurons and these interactions are under the

control of feedback cortico-thalamic connections [Jones 2012]. These strong cortico-geniculate feedback connections might explain why parallel retino-thalamo-cortical pathways are highly adaptive, dynamical systems [Mumford 1991, Cudeiro 2006, Briggs 2008]. In line with the computational constraints discussed before, both centre-surround interactions and feedback modulation can shape the dynamical properties of cortical inputs, maintaining the temporal precision of thalamic firing patterns during natural vision [Andolina 2007].

Overall, recent sub-cortical studies give us three main insights. First, we should not oversimplify the amount of processing done before visual inputs reach the cortex and we must instead consider that the retinal code is already highly structured, sparse and precise. Thus, we should consider how cortex takes advantage of these properties when processing naturalistic images. Second, some of the computational and mechanistic rules designed for predictive-coding or feature extraction can be much more generic than previously thought and the retina-LGN processing hierarchy may become again a rich source of inspiration for computer vision. Third, the exact implementation (what is being done and where) may be not so important as it varies from one species to another but the cascade of basic computational steps may be an important principle to retain from biological vision.

Functional and anatomical hierarchies are not always identical. The deep cortical hierarchy depicted in Fig. 7.1(b) is primarily based on gross anatomical connectivity rules [Zeki 1993]. Its functional counterpart is the increasing complexity of local processing and information content of expert cells as we go deeper along the anatomical hierarchy. There is however a flaw in attributing the functional hierarchy directly to its anatomical counterpart. The complexity of visual processing does increase from striate to extra-striate and associative cortices, but this is not attributable only to feedforward convergence. A quick glance at the actual cortical connectivity pattern in non-human primates would be sufficient to eradicate this textbook view of how the visual brain works [Hegdé 2007a, Markov 2013].

For example, a classical view is that the primary visual cortex represents luminance-based edges whereas higher-order image properties such as illusory contours are encoded at the next processing stages along the ventral path (e.g., areas V2 and V4) [Peterhans 1991]. Recent studies have shown however that illusory contours, as well as border ownerships can also be represented in macaque area V1 [Zhou 2000, Lee 2001]. Moreover, multiple binocular and monocular depth cues can be used to reconstruct occluded surfaces in area V1 [Sugita 1999]. Thus, the hierarchy of shape representation appears nowadays more opaque than previously thought [Hegde 2007b] and many evidences indicate that the intricate connectivity within and between early visual areas is decisive for of the emergence of figure-ground segmentation and proto-objects representations [Piech 2013, von der Heydt 2015]. Another strong example is visual motion processing. The classical feedforward framework proposes that MT cells (and not V1 cells) are true speed-tuned units. It has been thought for decades that V1 cells cannot encode the speed of a moving pattern independently of its spatiotemporal frequencies content [Rodman 1987]. However, recent studies have shown that there are V1 complex cells which are speed tuned [Priebe 2006]. The differences between V1 and MT regarding speed coding are more consistent with a distributed representation where slow speeds are represented in V1 and high speeds in area MT rather than a pure, serial processing.

Decoding visual motion information at multiple scales for elaborating a coherent motion percept must therefore imply a large-scale cortical network of densely recurrently interconnected areas. Such network can extend to cortical areas along the ventral stream in order to integrate together form and complex global motion inputs [Zhuo 2003, Hedges 2011]. One final example concerns the temporal dynamics of visual processing. The temporal hierarchy is not a carbon copy of the anatomical hierarchy depicted by Felleman and Van Essen. The onset of a visual stimulus triggers fast and slow waves of activation travelling throughout the different cortical areas. The fast activation in particular by-passes several major steps along both dorsal and ventral pathways to reach frontal areas even before area V2 is fully activated (for a review, see [Lamme 2000]). Moreover, different time scales of visual processing emerge from both the feedforward hierarchy of cortical areas but also from the long-range connectivity motifs and the dense recurrent connectivity of local sub-networks [Chaudhuri 2015]. Such rich repertoire of temporal time windows, ranging from fast, transient responses in primary visual cortex to persistent activity in association areas, is critical for implementing a series of complex cognitive tasks from low-level processing to decision-making.

These three different examples highlight the fact that a more complex view of the functional hierarchy is emerging. The dynamics of biological vision results from the interactions between different cortical streams operating at different speeds but also relies on a dense network of intra-cortical and inter-cortical (e.g., feedback) connections. Designing better vision algorithms could be inspired by this recurrent architecture where different spatial and temporal scales can be mixed to represent visual motion or complex patterns with both high reliability and high resolution.

Dorsal/ventral separation is an over-simplification. A strong limitation of grounding a theoretical framework of sensory processing upon anatomical data is that the complexity of connectivity patterns must lead to undesired simplifications in order to build a coherent view of the system. Moreover, it escapes the complexity of the dynamical functional interactions between areas or cognitive sub-networks. A good example of such bias is the classical dorsal/ventral separation. First, interactions between parallel streams can be tracked down to the primary visual cortex where a detailed analysis of the layer 4 connectivity have shown that both Magno and Parvocellular signals can be intermixed and propagated to areas V2 and V3 and, therefore the subsequent ventral stream [Yabuta 292]. Such a mixing of M- and P-like signals could explain why fast and coarse visual signals can rapidly tune the most ventral areas along the temporal cortex and therefore shape face recognition mechanisms [Giese 2003]. Second, motion psychophysics has demonstrated a strong influence of form signals onto local motion analysis and motion integration [Mather 2012]. These interactions have been shown to occur at different levels of the two parallel hierarchies, from primary visual cortex to the superior temporal sulcus and the parietal cortex [Orban 2008]. These interactions provide many computational advantages used by the visual motion system to resolve motion ambiguities, interpolate occluded information, segment the optical flow or recover the 3D structure of objects. Third, there are strong interactions between color and motion information, through mutual interactions between cortical areas V4 and MT [Thiele 2001a]. It is interesting to note that these two particular areas were previously attributed to the ventral and dorsal pathways, respectively [Livingstone 1988, DeYoe 1988]. Such

strict dichotomy is outdated as both V4 and MT areas interact to extract and mix these two dimensions of visual information.

These interactions are only a few examples to be mentioned here to highlight the needs of a more realistic and dynamical model of biological visual processing. If the coarse division between ventral and dorsal streams remains valid, a closer look at these functional interactions highlight the existence of multiple links, occurring at many levels along the hierarchy. Each stream is traversed by successive waves of fast/coarse and slow/precise signals so that visual representations are gradually shaped [Roelfsema 2005]. It is now timely to consider the intricate networks of intra and inter-cortical interactions to capture the dynamics of biological vision. Clearly, a new theoretical perspective on the cortical functional architecture would be highly beneficial to both biological and artificial vision research.

A hierarchy embedded within a dynamical recurrent system. We have already mentioned that spatial and temporal hierarchies do not necessarily coincide as information flows can bypass some cortical areas through fast cortico-cortical connections. This observation led to the idea that fast inputs carried by the Magnocellular stream can travel quickly across the cortical networks to shape each processing stage before it is reached by the fine-grain information carried by the Parvocellular retino-thalamo-cortical pathway. Such dynamics are consistent with the feedforward deep hierarchy and are used by several computational models to explain fast, automatic pattern recognition [Rousselet 2004, Thorpe 2009].

Several other properties of visual processing are more difficult to reconcile with the feedforward hierarchy. Visual scenes are crowded and it is not possible to process every of its details, Moreover, visual inputs are often highly ambiguous and can lead to different interpretations, as evidenced by perceptual multi-stability. Several studies have proposed that the highly recurrent connectivity motif of the primate visual system plays a crucial role in these processing. At the theoretical level, several authors recently resurrected the idea of a “reversed hierarchy" where high-level signals are back-propagated to the earliest visual areas in order to link low-level visual processing, high resolution representation and cognitive information [Bullier 2001, Hochstein 2002, Ahissar 2004, Gur 2015]. Interestingly, this idea was originally proposed more than three decades before by Peter Milner in the context of visual shape recognition [Milner 1974] and had then quickly diffused to the computer vision research leading to novel algorithms for top-down modulation, attention and scene parsing (e.g., [Fukushima 1987, Tsotsos 1993, Tsotsos 1995]). At the computational level, in [Lee 2003] the authors reconsidered the hierarchical framework by proposing that concatenated feedforward/feedback loops in the cortex could serve to integrate top-down prior knowledge with bottom-up observations. This architecture generates a cascade of optimal inference along the hierarchy [Roelfsema 2000, Lee 2003, Rousselet 2004, Thorpe 2009]. Several computational models have used such recurrent computation for surface motion integration [Bayerl 2004, Tlapale 2010, Perrinet 2012], contour tracing [Brosch 2015a] or figure-ground segmentation [Roelfsema 2002].

Empirical evidence for a role of feedback has long been difficult to gather in support to these theories. It was thus difficult to identify the constraints of top-down modulations that are known to play a major role in the processing of complex visual inputs, through selective attention, prior knowledge or action-related inter-

nal signals. However, new experimental approaches begin to give a better picture of their role and their dynamics. For instance, selective inactivation studies have begun to dissect the role of feedback signals in context-modulation of primate LGN and V1 neurons [Cudeiro 2006]. The emergence of genetically-encoded optogenetic probes targeting the feedback pathways in mice cortex opens a new era of intense research about the role of feedforward and feedback circuits [Luo 2008, Issacson 2011]. Overall, early visual processing appears now to be strongly influenced by different top-down signals about attention, working memory or even reward mechanisms, just to mention. These new empirical studies pave the way for a more realistic perspective on visual perception where both sensory inputs and brain states must be taken into account when, for example, modelling figure-ground segmentation, object segregation and target selection (see [Lamme 2000, Squire 2013, Kafaligonul 2015] for recent reviews).

The role of attention is illustrative of this recent trend. Mechanisms of bottom-up and top-modulation attentional modulations in primates have been largely investigated over the last three decades. Spatial and feature-based attentional signals have been shown to selectively modulate the sensitivity of visual responses even in the earliest visual areas [Motter 1993, Reynolds 2000]. These works have been a vivid source of inspiration for computer vision in searching for a solution to the problems of feature selection, information routing and task-specific attentional bias (see [Itti 2001, Tsotsos 2011]), as illustrated for instance by the Selective Tuning algorithm of Tsotsos and collaborators [Tsotsos 1995]. More recent work in non-human primates has shown that attention can also affect the tuning of individual neurons [Ibos 2014]. It also becomes evident that one needs to consider the effects of attention on population dynamics and the efficiency of neural coding (e.g., by decreasing noise correlation [Cohen 2009]). Intensive empirical work is now targeting the respective contributions of the frontal (e.g., task-dependency) and parietal (e.g., saliency maps) networks in the control of attention and its coupling with other cognitive processes such as reward learning or working memory (see [Buschman 2015] for a recent review). These empirical studies led to several computational models of attention (see [Tsotsos 2011, Tsotsos 2015, Bylinskii 2015] for recent reviews) based on generic computations (e.g., divisive normalisation [Reynolds 2009], synchrony [Fries 2005] or feedback-feedforward interactions [Khorsand 2015]). Nowadays, attention appears to be a highly dynamical, rapidly changing processing that recruits a highly flexible cortical network depending on behavioural demands and in strong interactions with other cognitive networks.

The role of lateral connectivity in information diffusion. The processing of a local feature is always influenced by its immediate surrounding in the image. Feedback is one potential mechanisms for implementing context-dependent processing but its spatial scale is rather large, corresponding to far-surround modulation [Angelucci 2006]. Visual cortical areas, and in particular area V1, are characterised by dense short- and long-range intra-cortical interactions. Short-range connectivities are involved in proximal centre-surround interactions and their dynamics fits with contextual modulation of local visual processing [Reynaud 2012]. This connectivity pattern has been overly simplified as overlapping, circular excitatory and inhibitory areas of the non-classical receptive field. In area V1, these subpopulations were described as being tuned for orthogonal orientations correspond-

ing to excitatory input from iso-oriented domains and inhibitory input from cross-oriented ones. In higher areas, similar simple schemes have been proposed, such as the opposite direction tuning of center and surround areas of MT and MST receptive fields [Born 2005]. Lastly, these surround inputs have been proposed to implement generic neural computations such as normalisation or gain control [Carandini 2011].

From the recent literature, a more complex picture of centre-surround interactions has emerged where non-classical receptive fields are highly diverse in terms of shapes or features selectivity [Xiao 1995, Cavanaugh 2002, Webb 2003]. Such diversity would result from complex connectivity patterns where neurons tuned for different features (e.g., orientation, direction, spatial frequency) can be dynamically interconnected. For example, in area V1, the connectivity pattern becomes less and less specific with farther distances from the recording sites. Moreover, far away points in the image can also interact through the long-range interactions which have been demonstrated in area V1 of many species. Horizontal connections extend over millimetres of cortex and propagate activity at a much lower speed than feedforward and feedback connections [Bullier 2001]. The functional role of these long-range connections is still unclear. They most probably support the waves of activity that travel across the V1 cortex either spontaneously or in response to a visual input [Sato 2012, Muller 2014]. They can also implement the spread of cortical activity underlying contrast normalisation [Reynaud 2012], the spatial integration of motion and contour signals [Reynaud 2012, Gilad 2013] or the shaping of low-level percepts [Jancke 2004].

A neural code for vision? How is information encoded in neural systems is still highly disputed and an active field of theoretical and empirical research. Once again, visual information processing has been largely used to decipher the neural coding principles and its application for computer sciences. The earliest studies on neuronal responses to visual stimuli have suggested that information is encoded in the mean firing rate of individual cells and its gradual change with visual input properties. For instance cells in V1 labelled as feature detectors are classified based upon their best response selectivity (stimulus that invokes maximal firing of the neuron) and several non-linear properties such gain control or context modulations which usually varied smoothly with respect to few attributes such as orientation contrast and velocity, leading to the development of tuning curves and receptive field doctrine. Spiking and mean-field models of visual processing are based on these principles.

Aside of from changes in mean firing rates, other interesting features of neural coding is the temporal signature of neural responses and the temporal coherence of activity between ensembles of cells, providing an additional potential dimension for specific linking, or grouping, distant and different features [von der Malsburg 1981, Von der Malsburg 1999, Singer 1999]. In networks of coupled neuronal assemblies, associations of related sensory features are found to induce oscillatory activities in a stimulus-induced fashion [Eckhorn 1990]. The establishment of a temporal coherence has been suggested to solve the so-called binding problem of task-relevant features through synchronization of neuronal discharge patterns in addition to the structural patterns of linking pattern [Engel 2001]. Such synchronizations might even operate over different areas and therefore seems to support rapid formations of neuronal groups and functional subnetworks and routing signals [Fries 2005, Buschman 2015]. However, the view that temporal oscillatory

states might define a key element of feature coding and grouping has been challenged by different studies and the exact contribution of these temporal aspects of neural codes is not yet fully elucidated (e.g., [Shadlen 1999] for a critical review). By consequences, only a few of bio-inspired and computer vision models rely on the temporal coding of information.

Although discussing the many facets of visual information coding is far beyond the scope of this review, one needs to briefly recap some key properties of neural coding in terms of tuning functions. Representations based on the tuning functions can be basis for the synergistic approach advocated in this thesis. Neurons are tuned to one or several features, i.e., exhibiting a strong response when stimuli contains a preferred feature such as local luminance-defined edges or proto-objects and low or no response when such features are absent. As a result, neural feature encoding is sparse, distributed over populations (see [Pouget 2013, Shamir 2014] and highly reliable [Perrinet 2015] at the same time. Moreover, these coding properties emerge from the different connectivity rules introduced above. The tuning functions of individual cells are very broad such that high behavioural performances observed empirically can be achieved only from some nonlinear or probabilistic decoding of population activities [Pouget 2013]. This could also imply that visual information could be represented within distributed population codes rather than grand-mother cells [Pouget 2003, Lehky 2013]. Tuning functions are dynamical: they can be sharpened or shifted over time [Shapley 2003]. Neural representation could also be relying on spike timing and the temporal structure of the spiking patterns can carry additional information about the dynamics of transient events [Thorpe 2001, Perrinet 2004]. Overall, the visual system appears to use different types of codes, one advantage for representing high-dimension inputs [Rolls 2010].

7.2 Computational studies of biological vision

7.2.1 The Marr's three levels of analysis

At conceptual level, much of the current computational understanding of biological vision is based on the influential theoretical framework defined by David Marr [Marr 1982] and colleagues. Their key message was that complex systems, like brains or computers, must be studied and understood at three levels of description: the computational task carried out by the system resulting in the observable behaviour, the instance of the algorithm used by the system to solve the computational task and the implementation that is emboddied by a given system to execute the algorithm. Once a functional framework is defined, the computational and implementation problems can be distinguished, so that in principle a given solution can be embedded into different biological, or artificial physical systems. This approach has inspired many experimental and theoretical research in the field of vision [Granlund 1978, Hildreth 1987, Daugman 1988, Poggio 2012]. The cost of this clear distinction between levels of description is that many of the existing models have only a weak relationship with the actual architecture of the visual system or even with a specific algorithmic strategy used by biological systems. Such dichotomy contrasts with the growing evidence that understanding cortical algorithms and networks are deeply coupled [Hildreth 1987]. Human

perception would still act as a benchmark or a source of inspiring computational ideas for specific tasks (see [Andreopoulos 2013] for a good example about object recognition). But, the risk of ignoring the structure-function dilemma is that computational principles would drift away from biology, becoming more and more metaphorical as illustrated by the fate of the Gestalt theory. The bio-inspired research stream for both computer vision and robotics aims at reducing this fracture (e.g. [Petrou 2008, Hérault 2010, Frisby 2010, Cristobal 2015] for recent reviews).

7.2.2 From circuits to behaviours

A key milestone in computational neurosciences is to understand how neural circuits lead to animal behaviours. Carandini [Carandini 2012] argued that the gap between circuits and behaviour is too wide without the help of an intermediate level of description, just that of neuronal computation. But how can we escape from the dualism between computational algorithm and implementation as introduced by Marr's approach? The solution depicted in [Carandini 2012] is based on three principles. First, some levels of description might not be useful to understand functional problems. In particular sub cellular and network levels are decoupled. Second, the level of neuronal computation can be divided into building blocks forming a core set of canonical neural computations such as linear filtering, divisive normalisation, recurrent amplification, coincidence detection, cognitive maps and so on. These standard neural computations are widespread across sensory systems [Fregnac 2015]. Third, these canonical computations occur in the activity of individual neurons and especially of population of neurons. In many instances, they can be related to stereotyped circuits such as feedforward inhibition, recurrent excitation-inhibition or the canonical cortical microcircuit for signal amplification (see [Sheperd 2010] for a series of reviews). Thus, understanding the computations carried out at the level of individual neurons and neural populations would be the key for unlocking the algorithmic strategies used by neural systems. This solution appears to be essential to capture both the dynamics and the versatility of biological vision. With such a perspective, computational vision would regain its critical role when mapping circuits to behaviours and could rejuvenate the interest in the field of computer vision not only by highlighting the limits of existing algorithms or hardware but also by providing new ideas. At this cost, visual and computational neurosciences would be again a source of inspiration for computer vision. To illustrate this joint venture, Figure 7.2 illustrates the relationships between the different functional and anatomical scales of cortical processing and their mapping with the three computational problems encountered with designing any artificial systems:how, what and why.

7.2.3 Neural constraints for functional tasks

Biological systems exist to solve functional tasks so that an organism can survive. Considering the existing constraints, many biologists consider the brain as a "bag of tricks that passed evolutionary selection", even though some tricks can be usable in different systems or contexts. This biological perspective highlights the fact that understanding biological systems is tightly related to understanding the functional importance of the task at hands. For example, there is in the mouse retina a cell type able to detect small moving objects in the presence of a fea-

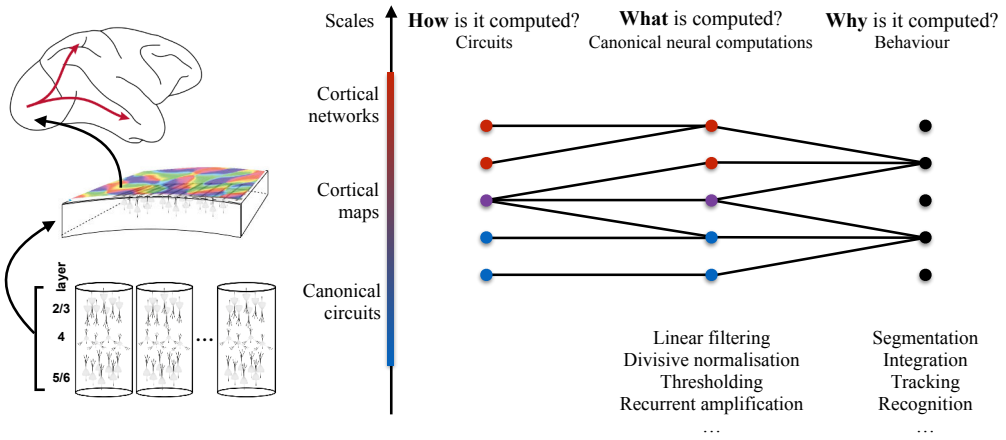


Figure 7.2: Between circuits and behaviour: rejuvenating the Marr approach. The nervous system can be described at different scales of organisation that can be mapped onto three computational problems: how, what and why. All three aspects involve a theoretical description rooted on anatomical, physiological and behaviour data. These different levels are organised around computational blocks that can be combined to solve a particular task.

tureless or stationary background. These neurons could serve as elementary detectors of potential predators arriving from the sky [Zhang 2012]. In the same vein, it has been recently found that output of retinal direction-selective cells are kept separated from the other retino-thalamo-cortical pathways to directly influence specific target neurons in mouse V1 [Cruz-Martin 2014]. These two very specific mechanisms illustrate how evolution can shape nervous systems. Computation and architecture are intrinsically coupled to find an optimal solution. This could be taken as an argument for ignoring neural implementations when building generic artificial systems. However, there are also evidence that evolution has selected neural microcircuits implementing generic computations such as divisive normalisation. These neural computations have been shown to play a key role in the emergence of low-level neuronal selectivities. For example divisive normalisation has been a powerful explanation for many aspects of visual perception, from low-level gain control or attention [Reynolds 2009, Carandini 2011]. The role of feedforward-feedback connectivity rules of canonical microcircuits in predictive coding have been also identified [Bastos 2012] and applied in the context of visual motion processing [Dimova 2009]. These examples are extrema lying on the continuum of biological structure-function solutions, from the more specific to the more generic. This diversity stresses the needs to clarify the functional context of the different computational rules and their performance dynamics so that fruitful comparisons can be made between living and artificial systems. This can lead to a clarification about which knowledge from biology is useful for computer vision.

Lastly, these computational building blocks are embedded into a living organism and low-to-high vision levels are constantly interacting with many other aspects of animal cognition [Vetter 2014]. For example, the way an object is examined (i.e., the way its image is processed) depends on its behavioural context, whether it is going to be manipulated or only scrutinised to identify it. A single face can be analysed in different ways depending upon the social or emotional context. Thus, we

must consider these contextual influence of "why" a task is being carried out when integrating information (and data) from biology [Willems 2011]. All these above observations stress the difficulty of understanding biological vision as an highly adapted, plastic and versatile cognitive system where circuits and computation are like Janus face. However, as described above for recurrent systems, understanding the neural dynamics of versatile top-down modulation can inspired artificial systems about how different belief states can be integrated together within the low-level visual representations.

7.2.4 Matching connectivity rules with computational problems

In Sec. 7.1, we have given a brief glimpse of the enormous literature on the intricate networks underlying biological vision. Focusing on primate low-level vision, we have illustrated both the richness, the spatial and temporal heterogeneity and the versatility of these connections. We illustrate them in Fig. 7.3 for a simple case, the segmentation of two moving surfaces. Figure 7.3(a) sketches the main cortical stages needed for a minimal model of surface segmentation [Orban 2008, Tlapale 2010]. Local visual information is transmitted upstream through the retinotopicaly-organized feedforward projections. In the classical scheme, V1 is seen as a router filtering and sending the relevant information along the ventral (V2, V4) or dorsal (MT, MST) pathways [Kruger 2013]. We discussed above how information flows also backward within each pathway as well as across pathways, as illustrated by connections between V2/V4 and MT in Fig. 7.3) [Markov 2014]. One consequence of these cross-over is that MT neurons are able to use both motion and color information [Thiele 2001a]. We have also highlighted that area V1 endorses a more active role where the thalamo-cortical feedforward inputs and the multiple feedback signals interact to implement contextual modulations over different spatial and temporal scales using generic neural computations such surround suppression, spatio-temporal normalisation and input selection. These local computations are modulated by short and long-range intra-cortical interactions such as visual features located far from the non-classical receptive field (or along a trajectory) can influence them [Angelucci 2003]. Each cortical stage implements these interactions although with different spatial and temporal windows and through different visual feature dimensions. In Fig. 7.3, these interactions are illustrated within two (orientation and position) of the many cortical maps founds in both primary and extra-striate visual areas. At the single neuron level, these intricate networks result in a large diversity of receptive field structures and in complex, dynamical non-linearities. It is now possible to collect physiological signatures of these networks at multiple scales, from single neurons to local networks and networks-of-networks such that connectivity patterns can be dissected out. In the near future, it will become possible to manipulate specific cell subtype and therefore change the functional role and the weight of these different connectivities.

How these connectivity patterns would relate to information processing? In Fig. 7.3(b) as an example, we sketch the key computational steps underlying moving surface segmentation [Braddick 1993]. Traditionally, each computational step has been attributed to a particular area and to a specific type of receptive fields. For instance, local motion computation is done at the level of the small receptive fields of V1 neurons. Motion boundary detectors have been found in area V2 while different

subpopulation of MT and MST neurons are responsible for motion integration at multiple scales (see Sec. 7.3.3 for references). However, each of these receptive field types are highly context-dependent, as expected from the dense interactions between all these areas. Matching the complex connectivity patterns illustrated in Fig. 7.3(a) with the computational dynamics illustrated in Fig. 7.3(b) is one of the major challenges in computational neurosciences [Fregnac 2015]. But it could also be a fruitful source of inspiration for computer vision if we were able to draw the rules and numbers by which the visual system is organised at different scales. So far, only a few computational studies have taken into account this richness and its ability to adaptively encode and predict sensory inputs from natural scenes (e.g., [Beck 2010, Bouecke 2011, Tlapale 2011b]. The goal of this review is to map such recurrent connectivity rules with the computational blocks and their dynamics. Thus, in Sec. 7.3 (see also Tables 7.2 and 7.1), we will recap some key papers from the biological vision literature in a task centric manner in order to show how critical information gathered at different scales and different context can be used to design innovative and performing algorithms.

In the context of the long-lasting debate about the precise relationships between structures and functions, we shall briefly mention the recent attempts to derive deeper insight about the processing hierarchy along the cortical ventral pathway. It has been suggested that deep convolutional neural networks (DCNNs) provide a potential framework for modelling biological vision. A directly related question is degree of similarity between the learning process implemented over several hierarchies in order to build feature layers of different selectivities with the cellular functional properties that have been identified in different cortical areas [Kriegeskorte 2015]. One proposal to generate predictive models of visual cortical function along the ventral path utilises a goal-driven approach to deep learning [Yamins 2016]. In a nutshell, such an approach optimises network parameters regarding performance on a task that is behaviourally relevant and then compares the resulting network(s) against neural data. As emphasised here, a key element in such a structural learning approach is to define the task-level properly and then map principled operations of the system onto the structure of the system. In addition, several parameters of deep networks are usually defined by hand, such a the number of layers or the number of feature maps within a layer. There have been recent proposals to optimise these automatically, e.g., by extensive searching or using genetic algorithms [Pinto 2009, Bergstra 2013].

7.2.5 Testing biologically-inspired models against both natural and computer vision

The dynamics of the biological visual systems have been probed at many different levels, from the psychophysical estimation of perceptual or behavioural performance to the physiological examination of neuronal and circuits properties. This diversity has led to a fragmentation of computational models, each targeting a specific set of experimental conditions, stimuli or responses.

Let consider visual motion processing in order to illustrate our point. When both neurally and psychophysically motivated models have been developed for a specific task such as motion integration for instance, they have been tested using a limited set of non-naturalistic inputs such as moving bars, gratings and plaid

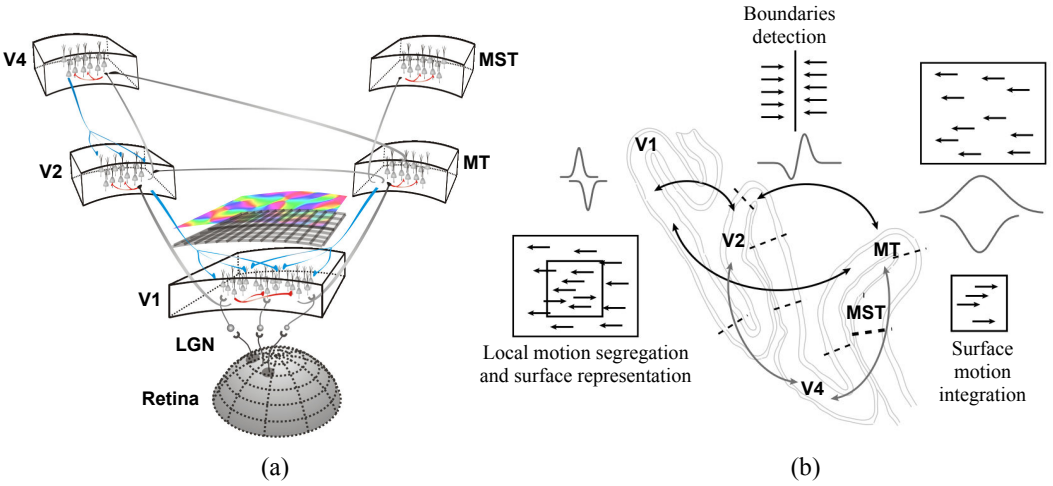


Figure 7.3: Matching multi-scale connectivity rules and computational problems for the segmentation of two moving surfaces. (a) A schematic view of the early visual stages with their different connectivity patterns: feedforward (grey), feedback (blue) and lateral (red). (b) A sketch of the problem of moving surface segmentation and its potential implementation in the primate visual cortex. The key processing elements are illustrated as computational problems (e.g., local segregation, surface cues, motion boundaries, motion integration) and corresponding receptive field structures. These receptive fields are highly adaptive and reconfigurable, thanks to the dense interconnections between the different stages/areas

patterns (e.g., [Nowlan 1994, Rust 2006]). These models formalise empirical laws that can explain either the perceived direction or the emergence of neuronal global motion direction preference. However, these models are hardly translated to velocity estimations in naturalistic motion stimuli since they do not handle scenarios such as lack of reliable cues or extended motion boundaries. By consequence, these models are very specific and not applicable directly to process generic motion stimuli. To overcome this limitation, a few extended computational models have been proposed that can cope with a broader range of inputs. These computational models handle a variety of complex motion inputs [Grossberg 2001, Tlapale 2010] but the specific algorithms have been tuned to recover coarse attributes of global motion estimation such as the overall perceived direction or the population neuronal dynamics. Such tuning strongly limits their ability to solve tasks such as dense optical flow estimation. Still, their computational principles can be used as building blocks to develop extended algorithms that can handle naturalistic inputs [Perrone 2012, Solari 2015]. Moreover, they can be evaluated against standard computer vision benchmarks [Baker 2011, Butler 2012]. What is still missing are detailed physiological and psychophysical data collected with complex scenarios such as natural or naturalistic images in order to be able to further constrain these models.

A lesson to be taken from the above example is that a successful synergistic approach between artificial and natural vision should first establish a common set of naturalistic inputs against which both bio-inspired and computer vision models can be benchmarked and compared. This step is indeed critical for identifying scenarios in which biological vision systems deviate with respect to the definition

adopted by the computer vision. On the other side, state-of-the-art computer vision algorithms shall also be evaluated relative to human perception performance for the class of stimuli widely used in psychophysics. For the three illustrative tasks to be discussed below, we will show the interest of common benchmarks for comparing biological and computer vision solutions.

7.2.6 Task-based versus general purpose vision systems

Several objections can be raised to question the need for a synergy between natural and biological vision. A first objection is that biological and artificial systems could serve different aims. In particular, the major aim of biological vision studies is to understand the behaviours and properties of a general purpose visual system that could subserve different types of perceptions or actions. This generic, encapsulated visual processing machine can then be linked with other cognitive systems in an adaptive and flexible way (see [Pylyshyn 1999, Tsotsos 2011] for example). By contrast, computer vision approaches are more focused on developing task specific solutions, with an ever growing efficiency thank to advances in algorithms (e.g., [LeCun 2015, Mnih 2015]) supported by growing computing power. A second objection is that the brain might not use the same general-purpose (Euclidean) description of the world that Marr postulated [Warren 2012]. Thus perception may not use the same set of low-level descriptors as computer vision, dooming the search for common early algorithms. A third, more technical objection is related to the low performance of most (if not all) current bio-inspired vision algorithms when solving a specific task (e.g., face recognition) when compared to state-of-the-art computer vision solutions. Moreover, bio-inspired models are still too often based on over-simplistic inputs and conditions and not sufficiently challenged with high-dimension inputs such as complex natural scenes or movies. Finally, artificial systems can solve a particular task with a greater efficiency than human vision for instance, challenging the need for bio-inspiration.

These objections question the interest of grounding computer vision solution on biology. Still, many other researchers have argued that biology can help recasting ill-based problems and showing us to ask the right questions and identifying the right constraints [Zucker 1981, Tsotsos 2014]. Moreover, to mention one recent example, perceptual studies can still identify feature configurations that cannot be used by current models of object recognition and thus reframing the theoretical problems to be solved to match human performance [Ullman 2016]. Finally, recent advances in computational neurosciences has identified generic computational modules that can be used to solve several different perceptual problems such as object recognition, visual motion analysis or scene segmentation, just to mention a few (e.g. [Carandini 2011, Cox 2014, Fregnac 2015]). Thus, understanding task-specialised subsystems by building and testing them remains a crucial step to unveil the computational properties of building blocks that operate in largely unconstrained scene conditions and that could later be integrated into larger systems demonstrating enhanced flexibility, default-resistance or learning capabilities. Theoretical studies have identified several mathematical frameworks for modelling and simulating these computational solutions that could be inspiring for computer vision algorithms. Lastly, current limitations of existing bio-inspired models in terms of their performance will also be solved by scaling up and tuning them such that they pass the

traditional computer vision benchmarks.

We propose herein that the task level approach is still an efficient framework for this dialogue. Throughout the next sections, we will illustrate this standpoint with three particular examples: retinal image sensing, scene segmentation and optic flow computation. We will highlight some important novel constraints emerging from recent biological vision studies, how they have been modelled in computational vision and how they can lead to alternative solutions.

7.3 Solving vision tasks with a biological perspective

In the preceding sections, we have revisited some of the main features of biological vision and we have discussed the foundations of the current computational approaches of biological vision. A central idea is the functional importance of the task at hand when exploring or simulating the brain. Our hypothesis is that such a task centric approach would offer a natural framework to renew the synergy between biological and artificial vision. We have discussed several potential pitfalls of this task-based approach for both artificial and bio-inspired approaches. But we argue that such task-centric approach will escape the difficult, theoretical question of designing general-purpose vision systems for which no consensus is achieved so far in both biology and computer vision. Moreover, this approach allow us to benchmark the performance of computer and bio-inspired vision systems, an essential step for making progress in both fields. Thus, we believe that the task-based approach remains the most realistic and productive approach. The novel strategy based on bio-inspired generic computational blocks will however open the door for improving the scalability, the flexibility and the fault-tolerance of novel computer vision solutions. As already stated above, we decided to revisit three classical computer vision tasks from such a biological perspective: image sensing, scene segmentation and optical flow.[1] This choice was made in order to provide a balanced overview of recent biological vision studies about three illustrative stages of vision, from the sensory front-end to the ventral and dorsal cortical pathways. For these three tasks, there are a good set of multiple scales biological data and a solid set of modelling studies based on canonical neural computational modules. This enables us to compare these models with computer vision algorithms and to propose alternative strategies that could be further investigated. For the sake of clarity, each task will be discussed with the following framework:

Task definition. We start with a definition of the visual processing task of interest.

Core challenges. We summarise its physical, algorithmic or temporal constraints and how they impact the processing that should be carried on images or sequences of images.

Biological vision solution. We review biological facts about the neuronal dynamics and circuitry underlying the biological solutions for these tasks stressing

[1] See also, recent review articles addressing other tasks: object recognition [Andreopoulos 2013], visual attention [Tsotsos 2011, Tsotsos 2015], biological motion [Giese 2003].

the canonical computing elements being implemented in some recent computational models.

Comparison with computer vision solutions. We discuss some of the current approaches in computer vision to outline their limits and challenges. Contrasting these challenges with known mechanisms in biological vision would be to foresee which aspects are essential for computer vision and which ones are not.

Promising bio-inspired solutions. Based on this comparative analysis between computer and biological vision, we discuss recent modelling approaches in biological vision and we highlight novel ideas that we think are promising for future investigations in computer vision.

7.3.1 Sensing

Task definition. Sensing is the process of capturing patterns of light from the environment so that all the visual information that will be needed downstream to cater the computational/functional needs of the biological vision system could be faithfully extracted. This definition does not necessarily mean that its goal is to construct a veridical, pixel-based representation of the environment by passively transforming the light the sensor receives.

Core challenges. From a functional point of view, the process of sensing (i.e., transducing, transforming and transmitting) light patterns encounters multiple challenges because visual environments are highly cluttered, noisy and diverse. First, illumination levels can vary over several range of magnitudes. Second, image formation onto the sensor is sensitive to different sources of noise and distortions due to the optical properties of the eye. Third, transducing photons into electronic signals is constrained by the intrinsic dynamics of the photosensitive device, being either biological or artificial. Fourth, transmitting luminance levels on a pixel basis is highly inefficient. Therefore, information must be (pre-)processed so that only the most relevant and reliable features are extracted and transmitted upstream in order to overcome the limited bandpass properties of the optic nerve. At the end of all these different stages, the sensory representation of the external world must still be both energy and computationally very efficient. All these aforementioned aspects raise some fundamental questions that are highly relevant for both modelling biological vision and improving artificial systems.

Herein, we will focus on four main computational problems (what is computed) that are illustrative about how biological solutions can inspire a better design of computer vision algorithms. The first problem is called *adaptation* and explains how retinal processing is adapted to the huge local and global variations in luminance levels from natural images in order to maintain high visual sensitivity. The second problem is *feature extraction*. Retinal processing extracts information about the structure of the image rather than mere pixels. What are the most important features that sensors should extract and how they are extracted are pivotal questions that must be solved to sub-serve an optimal processing in downstream networks. Third is the *sparseness* of information coding. Since the amount of information that can be transmitted from the front-end sensor (the retina) to the central

processing unit (area V1) is very limited, a key question is to understand how spatial and temporal information can be optimally encoded, using context dependency and predictive coding. The last selected problem is called *precision* of the coding, in particular what is the temporal precision of the transmitted signals that would best represent the seaming-less sequence of images.

Biological vision solution. The retina is one of the most developed sensing devices [Gollisch 2010, Masland 2011, Masland 2012]. It transforms the incoming light into a set of electrical impulses, called spikes, which are sent asynchronously to higher level structures through the optic nerve. In mammals, it is sub-divided into five layers of cells (namely, photoreceptors, horizontal, bipolar, amacrine and ganglion cells) that forms a complex recurrent neural network with feedforward (from photoreceptors to ganglion cells), but also lateral (i.e., within bipolar and ganglion cells layers) and feedback connections. The complete connectomics of some invertebrate and vertebrate retinas now begin to be available [Marc 2013].

Regarding information processing, an humongous amount of studies have shown that the mammalian retina can tackle the four challenges introduced above using *adaptation*, *feature detection*, *sparse coding* and *temporal precision* [Kastner 2014]. Note that *feature detection* should be understood as "feature encoding" in the sense that there is non decision making involved. Concerning *adaptation*, it is a crucial step, since retinas must maintain high contrast sensitivity over a very broad range of luminance, from starlight to direct sunlight. Adaptation is both global through neuromodulatory feedback loops and local through adaptive gain control mechanisms so that retinal networks can be adapted to the whole scene illuminance level while maintaining high contrast sensitivity in different regions of the image, despite their considerable differences in luminance [Shapley 1984, Demb 2008, Thoreson 2012].

It has long been known that retinal ganglion cells extract local luminance profiles. However, we have now a more complex view of retinal form processing. The retina of higher mammals sample each point in the images with about 20 distinct ganglion cells [Masland 2011, Masland 2012] associated to different *features*. This is best illustrated in Fig. 7.4, showing how the retina can gather information about the structure of the visual scene with four example cell types tilling the image. They differ one from the others by the size of their receptive field and their spatial and temporal selectivities. These spatiotemporal differences are related to the different sub-populations of ganglion cells which have been identified. Parvocellular (P) cells are the most numerous are the P-cells (80%). They have a small receptive size and a slow response time resulting in a high spatial resolution and a low temporal sensitivity. They process information about color and details. Magnocellular cells have a large receptive field and a low response time resulting in a high temporal resolution and a low spatial sensitivity, and can therefore convey information about visual motion [Shapley 1990]. Thus visual information is split into parallel stream extracting different domains of the image spatiotemporal frequency space. This was taken at a first evidence for feature extractions at retinal level. More recent studies have shown that, in many species, retinal networks are much smarter than originally thought. In particular, they can extract more complex features such as basic static or moving shapes and can predict incoming events, or adapt to temporal changes of events, thus exhibiting some of the major signatures of predictive coding [Gollisch 2010, Masland 2011, Masland 2012].

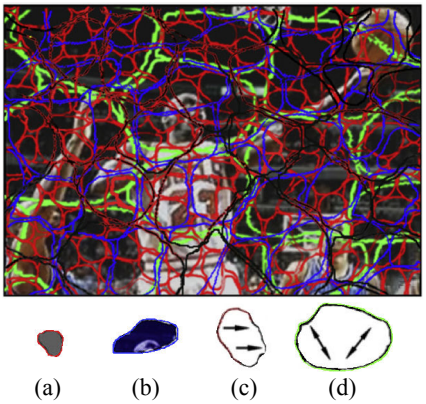


Figure 7.4: How retinal ganglion cells tile a scene extracting a variety of features. This illustrates the tiling of space of a subset of four cell types. Each tile covers completely the visual image independently from other types. The four cell types shown here correspond to (a) cell with small receptive fields and center-surround characteristics extracting intensity contrasts, (b) color coded cells, (c) motion direction selective cells with a relatively large receptive field, (d) cells with large receptive fields reporting that something is moving (adapted from [Masland 2012], with permissions).

A striking aspect of retinal output is its *high temporal precision* and *sparseness*. Massive in vitro recordings provide spiking patterns collected from large neuronal assemblies so that it becomes possible to decipher the retinal encoding of complex images [Pillow 2008]. Modelling the spiking output of the ganglion cell populations have shown high temporal precision of the spike trains and a strong reliability across trials. These coding properties are essential for upstream processing what will extract higher order features but also will have to maintain such high precision. In brief, the retina appears to be a dense neural network where specific sub-populations adaptively extract local information in a context-dependent manner in order to produce an output that is both adaptive, sparse, over complete and of high temporal precision.

Another aspect of retinal coding is its space-varying resolution. A high-resolution sampling zone appears in the fovea while the periphery looses spatial detail. The retinotopic mapping of receptors into the cortical representation can be characterized formally by a non-linear conformal mapping operation. Different closed-form models have been proposed which share the property that the retinal image is sampled in a space-variant fashion using a topological transformation of the retinal image into the cortex. The smooth variation of central into peripheral vision may directly support a mechanism of space-variant vision. Such active processing mechanism not only significantly reduces the amount of data (particularly with a high rate of peripheral compression) but may also support computational mechanisms, such as symmetry and motion detection.

There is a large, and expanding body of literature proposing models of retinal processing. We attempted to classify them and isolated three main classes of models. The first class regroups the linear-nonlinear-poisson (LNP) models [Odermatt 2012]. In its simplest form, a LNP model is a convolution with a spatio-temporal kernel followed by a static nonlinearity and stochastic (Poisson-like) mechanisms of spikes

generation. These functional model are widely used by experimentalists to characterise the cells that they record, map their receptive field and characterise their spatiotemporal feature selectivities [Chichilnisky 2001]. LNP models can simulate the spiking activity of ganglion cells (and of cortical cells) in response to synthetic or natural images [Carandini 2005] but they voluntarily ignore the neuronal mechanisms and the details of the inner retinal layers that transform the image into a continuous input to the ganglion cell (or any type of cell) stages. Moreover, they implement static non-linearities, ignoring many existing non-linearities. Applied to computer vision, they however provide some inspiring computational blocks for contrast enhancement, edge detection or texture filtering.

The second class of models has been developed to serve as a front-end for subsequent computer vision task. They provide bio-inspired modules for low level image processing. One interesting example is given by [Benoit 2010, Hérault 2010], where the model includes parvocellular and magnocellular pathways using different non-separable spatio-temporal filter that are optimal for form or motion detection.

The third class is based on detailed retinal models reproducing its circuitry, in order to predict the individual or collective responses measured at the ganglion cells level [Wohrer 2009, Lorach 2012]. Virtual Retina [Wohrer 2009] is one example of such spiking retina model. This models enables large scale simulations (up to 100,000 neurons) in reasonable processing times while keeping a strong biological plausibility. These models are expanded to explore several aspects of retinal image processing such as (i) understanding how to reproduce accurately the statistics of the spiking activity at the population level [Nasser 2013], (ii) reconciling connectomics and simple computational rules for visual motion detection [Kim 2014] and (iii) investigating how such canonical microcircuits can implement the different retinal processing modules cited above (feature extraction, predictive coding) [Gollisch 2010].

Comparison with computer vision solutions. Most computer vision systems are rooted on a sensing device based on CMOS technology to acquire images in a frame based manner. Each frame is obtained from sensors representing the environment as a set of pixels whose values indicate the intensity of light. Pixels pave homogeneously the image domain and their number defines the resolution of images. Dynamical inputs, corresponding to videos are represented as a set of frames, each one representing the environment at a different time, sampled at a constant time step defining the frame rate.

To make an analogy between the retina and typical image sensors, the dense pixels which respond slowly and capture high resolution color images are at best comparable to P-Cells in the retina. Traditionally in computer vision, the major technological breakthroughs for sensing devices have aimed at improving the density of the pixels, as best illustrated by the ever improving resolution of the images we capture daily with cameras. Focusing of how videos are captured, one can see that a dynamical input is not more that a series of images sampled at regular intervals. Significant progress have been achieved recently in improving the temporal resolution with advent of computational photography but at a very high computational cost [Liu 2014]. This kind of sensing for videos introduces a lot of limitations and the amount of data that has to be managed is high.

However, there are two main differences between the retina and a typical image sensor such as a camera. First, as stated above, the retina is not simply sending

an intensity information but it is already extracting features from the scene. Second, the retina asynchronously processes the incoming information, transforming it as a continuous succession of spikes at the level of ganglion cells, which mostly encode changes in the environment: retina is very active when intensity is changing, but its activity becomes quickly very low with a purely static stimulation. These observations show that the notion of representing static frames does not exist in biological vision, drastically reducing the amount of data that is required to represent temporally varying content.

Promising bio-inspired solutions. Analysing the sensing task from a biological perspective has potential for bringing new insights and solutions related to the four challenges outlined in this section. In terms of an ideal sensor, it is desired to have control over the acquisition of each pixel, thus allowing a robust adaptation to different parts of the scene. However, this is difficult to realize on the chip as it would mean independent triggers to each pixel, thus increasing the information transfer requirements on the sensor. In order to circumvent this problem, current CMOS sensors utilize a global clock trigger which fails us to give a handle on local adaptation, thus forcing a global strategy. This problem is tackled differently in biologically inspired sensors, by having local control loops in the form of event driven triggering rather than a global clock based drive. This helps the sensor to adapt better to local changes and avoids the need for external control signals. Also, since the acquisitions are to be rendered, sensory physiological knowledge could help in choosing good tradeoffs on sensor design. For example, the popular Bayer filter pattern has already been inspired by the physiological properties of retinal color sensing cells. With the advent of high dynamic range imaging devices, these properties are beginning to find interesting applications such as low range displays. This refers to the tone mapping problem. It is a necessary step to visualize high-dynamic range images on low-dynamic range displays, spanning up to two orders of magnitude. There is a large body of literature in this area on static images (see [Kung 2007, Bertalmío 2014] for reviews), with approaches which combine luminance adaptation and local contrast enhancement sometimes closely inspired from retinal principles, as in [Meylan 2007, Benoit 2009, Ferradans 2011, Muchungi 2012] just to cite a few. Recent developments concern video-tone mapping where a few approaches have been developed so far (see [Eilertsen 2013] for a review). We think it is for videos that the development of synergistic models of the retina is the most promising. Building on existing detailed retinal models such as the Virtual Retina [Wohrer 2009] (mixing filter-based processing, dynamical systems and spiking neuron models), the goal is to achieve a better characterization of retinal response dynamics which will have a direct application here.

The way that retina performs *feature detection* and encodes information in space and time has received relatively little attention so far from the computer vision community. In most cases, retina-based models rely on simple caricatures of the retina. The FREAK (Fast Retina Keypoint) descriptor [Alahi 2012] is one example where only the geometry and space-varying resolution has been exploited. In [Alahi 2012], the "cells" in the model are only doing some averaging of intensities inside their receptive field. This descriptor model was extended in [Hilario Gomez 2015] where ON and OFF cells were introduced using a linear-nonlinear (LN) model. This gives a slight gain of performance in a classification task, although it is still far from the

state-of-the-art. These descriptors could be improved in many ways, by taking into account the goal of the features detected by the 20 types of ganglion cells mentioned before. Here also the strategy is to build on existing retinal models. In this context, one can also mention the SIFT descriptor [Lowe 2001] which was also inspired by cortical computations. One needs to evaluate the functional implication at a task level of some retinal properties. Examples include the asymmetry between ON and OFF cells [Pandarinath 2010] and the irregular receptive field shapes [Liu 2009].

One question is whether we would still need inspiration from the retina to build new descriptors, given the power of machine learning methods that provides automatically some optimized features given an image database? What the FREAK-based models show is that it is not only about improving the filters. It is also about how the information is encoded. In particular, what is encoded in FREAK-based models is the relative difference between cell responses. Interestingly, this is exactly the same as the rank-order coding idea proposed as an efficient strategy to perform ultra-fast categorization [VanRullen 2002], and which has been reported in the retina [Portelli 2014]. This idea has been exploited for pattern recognition and used in many applications as demonstrated by the products developed by the company Spikenet (http://www.spikenet-technology.com). This means that the retina should serve as a source of inspiration not only to propose features, but more importantly, how it encodes these features at a population level.

The fact that the retinal output is *sparse* and has a *high temporal precision* conveys a major advantage to the visual system, since it has to deal with only a small amount of information. A promising bio-inspired solution is to develop frame-free methods, i.e., methods using sparse encoding of the visual information. This is now possible using event-based vision sensors where pixels autonomously communicate the change and grayscale events. The dynamic vision sensor (DVS) [Lichtsteiner 2008, Liu 2010] and the asynchronous time-based image sensor (ATIS) [Posch 2011] are two examples of such sensor using address-event representation (AER) circuits. The main principle is that pixels signal only significant events. More precisely, an event is sent when the log intensity has changed by some threshold amount since the last event (see Fig. 7.5). These sensors provide a sparse output corresponding to pixels that register a change in the scene, thus allowing extremely high temporal resolution to describe changes in the scene while discarding all the redundant information. Because the encoding is sparse, these sensors appear as a natural solution in real-time scenarios or when energy consumption is a constraint. Combined with what is known about retinal circuitry as in [Lorach 2012], they could provide a very efficient front-end for subsequent visual tasks, in the same spirit of former neuromorphic models of low-level processing as in [Benoit 2010, Hérault 2010]. They could also be used more directly as a way to represent visual scenes, abandoning the whole notion of a video that is composed of frame-sequences. This provides a new operative solution that can be used to revisit computer vision problems (see [Liu 2015] for a review). This field is rapidly emerging, with the motivation to develop approaches more efficient than the state-of-the-art. Some examples include tracking [Ni 2011], stereo [Rogister 2012], 3D pose estimation [Valeiras 2016], object recognition [Orchard 2015] and optical flow [Benosman 2011, Tschechne 2014b, Brosch 2015b, Giuliani 2016].

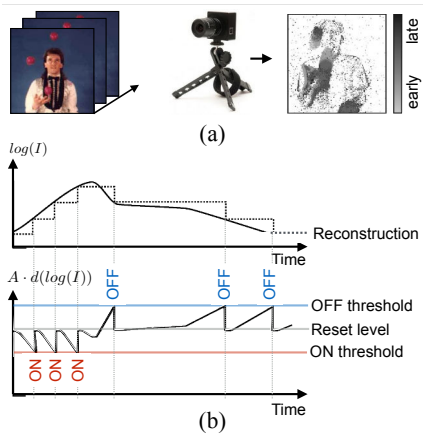


Figure 7.5: How DVS sensor generate spikes. (a) Example of a video with fast motions (a juggling scene). DVS camera and DVS output: Events are rendered using a grayscale colormap corresponding to events that were integrated over a brief time window (black = young, gray = old, white = no events). (b) DVS principle: Positive and negative changes are generated depending on the variations of $log(I)$ which are indicated as ON and OFF events along temporal axis (adapted from [Lichtsteiner 2008], with permissions).

7.3.2 Segmentation and figure-ground segregation

Task definition. The task of segmenting a visual scene is to generate a meaningful partitioning of the input feature representation into surface- or object-related components. The segregation of an input stimulus into prototypical parts, characteristic of surfaces or objects, is guided by a coherence or homogeneity property that region elements share. Homogeneities are defined upon feature domains such as color, motion, depth, statistics of luminance items (texture), or combinations of them [Pal 1993, Martin 2001]. The specificity of the behavioural task, e.g., grasping an object, distinguishing two object identities, or avoiding collisions during navigation, may influence the required detail of segmentation [Ballard 2000, Hayhoe 2005]. In order to do so, contextual information in terms of high-level knowledge representations can be exploited as well [Borenstein 2008]. In addition, the goal of segmentation might be extended in regard to eventually single out a target item, or object, from its background in order to recognise it or to track its motion.

Core challenges. The segmentation of a spatio-temporal visual image into regions that correspond to prototypical surfaces or objects faces several challenges which derive from distinct interrelated subject matters. The following themes refer to issues of *representation*. First, the feature domain or multiple domains need to be identified which constitute the coherence or homogeneity properties relevant for the segregation task. Feature combinations as well as the nested structure of their appearance of coherent surfaces or objects introduces apparent feature hierarchies [Koenderink 1984, Koenderink 2012]. Second, the segmentation process might focus on the analysis of homogeneities that constitute the coherent components within a region or, alternatively, on the discontinuities between regions of homogeneous appearances. Approaches belonging to the first group focus on the segregation

of parts into meaningful prototypical regions utilising an agglomeration (clustering) principle. Approaches belonging to the second group focus on the detection of discontinuous changes in feature space (along different dimensions) [Nothdurft 1991] and group them into contours and boundaries. Note that we make a distinction here to refer to a contour as a grouping of oriented edge or line contrast elements whereas a boundary already relates to a surface border in the scene. Regarding the boundaries of any segment, the segmentation task itself might incorporate an explicit assignment of a border ownership (BOwn) direction label which implies the separation of figural shape from background by a surface that occludes other scenic parts [Peterson 2008, Kogo 2013]. The variabilities in the image acquisition process caused by, e.g., illumination conditions, shape and texture distortions, might speak in favor of a boundary oriented process. On the other hand, the complexity of the background structure increases the effort to segregate a target object from the background, which argues in favour of region oriented mechanisms. It should be noted, however, that the region vs boundary distinction might not appear as binary as in the way outlined above. Considering real world scenes the space-time relationships of perceptual elements (defined over different levels of resolution) are often defined by statistically meaningful structural relations to determine segmentation homogeneities [Witkin 1983]. Here, an important distinction has been made between structure that might be influenced by meaning and primitive structure that is perceived even without a particular interpretation.

While the previous challenges were defined by representations, the following themes refer to the process characteristic of segmentation. First, the partitioning process may yield different results given changing view-points or different noise sources during the sensing process. Thus, segmentation imposes an inference problem that is mathematically ill-posed [Poggio 1985]. The challenge is how a reliability, or confidence, measure is defined that characterises meaningful decompositions relating to reasonable interpretations. To illustrate this, Fig. 7.6 shows segmentation results as drawn by different human observers. Second, figural configurations may impose different efforts for mechanisms of perceptual organisation to decide upon the segregation of an object from the background and/or the assignment of figure and ground direction of surface boundaries. A time dependence that correlates with the structural complexity of the background has in fact been observed to influence the temporal course needed in visual search tasks [Wolfe 2002].

Biological vision solution. Evidence from neuroscience suggests that the visual system uses segmentation strategies based on identifying discontinuities and grouping them into contours and boundaries. Such processes operate mainly in a feedforward fashion and automatic, utilising early and intermediate-level stages in visual cortex. In a nutshell, contrast and contour detection is quickly accomplished and is already represented at early stages in the visual cortical hierarchy, namely areas V1 and V2. The assignment of task-relevant segments happens to occur after a slight temporal delay and involves a recurrent flow of lateral and feedback processes [Roelfsema 2006, Scholte 2008, Roelfsema 2011].

The grouping of visual elements into contours appears to follow the Gestalt rules of perceptual organisation [Koffka 1935]. Grouping has also been studied in accordance to the ecological validity of such rules as they appear to be embedded in the statistics of natural scenes [Brunswik 1953]. Mechanisms that entail

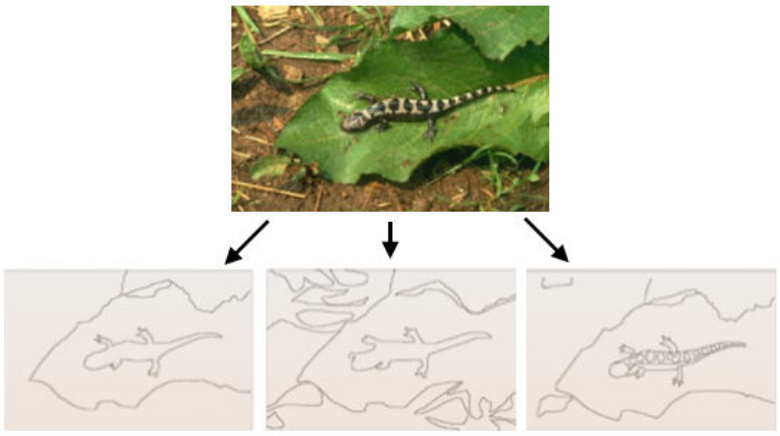


Figure 7.6: Example of possible segmentation results for a static image drawn by different human observers. Lower images shows segmentations happening at different levels of detail but consistent with each other (adapted from [Arbelaez 2011]).

contour groupings are implemented in the structure of supragranular horizontal connections in area V1 in which oriented cells preferentially contact like-oriented cells that are located along the orientation axes defined by a selected target neuron [Kapadia 1995, Bosking 1997]. Such long-range connections form the basis for the Gestalt concept of good continuation and might reflect the physiological substrate of the association field, a figure-eight shaped zone of facilitatory coupling of orientation selective input and perceptual integration into contour segments [Grossberg 1985, Field 1993, Geisler 2001]. Recent evidence suggests that the perceptual performance of visual contour grouping can be improved by mechanisms of perceptual learning [Li 2008]. Once contours have been formed they need to be labelled in accordance to their scene properties. In case of a surface partially occluding more distant scenic parts the *border ownership (BOwn) or surface belongingness* can be assigned to the boundary [Koffka 1935]. A neural correlate of such a mechanism has been identified at different cortical stages along the ventral pathway, such as V1, V2 and V4 areas [Zhou 2000, O'Herron 2011]. The dynamics of the generation of the BOwn signals may be explained by feedforward, recurrent lateral and feedback mechanisms (see [Williford 2013] for a review).

Such dynamical process of feedback, called re-entry [Edelman 1993], recursively links representations distributed over different levels. Mechanisms of lateral integration, although slower in processing speed, seem to further support intra-cortical grouping [Kapadia 1995, Kapadia 2000, Gilbert 2013]. In addition, surface segregation is reflected in a later temporal processing phase but is also evident in low levels of the cortical hierarchy, suggesting that recurrent processing between different cortical stages is involved in generating neural surface representations. Once boundary groupings are established surface-related mechanisms "paint", or tag, task-relevant elements within bounded regions. The feature dimensions used in such grouping operations are, e.g., local contour orientations defined by luminance contrasts, direction and speed of motion, color hue contrasts, or texture orientation gradients. As sketched above, counter-stream interactive signal flow [Ullman 1995] imposes a temporal signature on responses in which after a delay a late amplification signal serves to tag those local responses that belong to a region (surrounded by contrasts)

which has been selected as a figure [Lamme 1995] (see also [Roelfsema 2007]). The time course of the neuronal responses encoding invariance against different figural sizes argues for a dominant role of feedback signals when dynamically establishing the proper BOwn assignment. Grouping cells have been postulated that integrate (undirected) boundary signals over a given radius and enhance those configurations that define locally convex shape fragments. Such fragments are in turn enhanced via a recurrent feedback cycle so that closed shape representations can be established rapidly through the convexity in closed bounding contours [Zhou 2000]. Neural representations of localized features composed of multiple orientations may further influence this integration process, although this is not firmly established yet [Anzai 2007]. BOwn assignment serves as a prerequisite of figure-ground segregation. The temporal dynamics of cell responses at early cortical stages suggest that mechanisms exist that (i) decide about ownership direction and (ii) subsequently enhance regions (at the interior of the outline boundaries) by spreading a neural tagging, or labelling, signal that is initiated by the region boundary [Roelfsema 2002] (compare the discussion in [Williford 2013]). Such a late enhancement through response modulation of region components occurs for different features, such as oriented texture [Lamme 1999] or motion signals [Roelfsema 2007], and is mediated by recurrent processes of feedback from higher levels in the cortical hierarchy. It is, however, not clear whether a spreading process for region tagging is a basis for generating invariant neural surface representations in all cases. All experimental investigations have been conducted for input that leads to significant initial stimulus responses while structure-less homogeneous regions (e.g., a homogeneous coloured wall) may lead to void spaces in the neuronal representation that may not be filled explicitly by the cortical processing (compare the discussion in [Pessoa 1998]).

Yet another level of visual segmentation operates upon the initial grouping representations, those base groupings that happen to be processed effortlessly as outlined above. However, the analysis of complex relationships surpasses the capacities of the human visual processor which necessitates serial staging of some higher-level grouping and segmentation mechanisms to form incremental task-related groupings. In this mainly sequential operational mode visual routines establish properties and relations of particular scene items [Ullman 1984]. Elemental operations underlying such routines have been suggested, e.g., shifting the processing focus (related to attentional selection), indexing (to select a target location), coloring (to label homogeneous region elements), and boundary tracing (determining whether a contour is open or closed and items belonging to a continuous contour). For example, contour tracing is suggested to be realized by incremental grouping operations which propagate an enhancement of neural firing rates along the extent of the contour. Such a neural labelling signal is reflected in a late amplification in the temporal signature of neuronal responses. The amplification is delayed with respect to the stimulus onset time with increasing distances of the location along the perceptual entity [Jolicoeur 1986, Roelfsema 2011] (that is indexed by the fixation point at the end of the contour). This lead to the conclusion that such tracing is laterally propagated (via lateral or interative feedforward and feedback mechanisms), leading to a neural segmentation of the labelled items delineating feature items that belong to the same object or perceptual unit. Maintenance operations then interface such elemental operations into sequences to compose visual routines for solving more complex tasks, like in a sequential computer program. Such cognitive opera-

tions are implemented in cortex by networks of neurons that span several cortical areas [Roelfsema 2005]. The execution time of visual cortical routines reflects the sequential composition of such task-specific elemental neural operations tracing the signature of neural responses to a stimulus [Lamme 2000, Roelfsema 2005].

Comparison with computer vision solutions. Segmentation as an intermediate level process in computational vision is often characterised as one of agglomerating, or clustering, picture elements to arrive at an abstract description of the regions in a scene [Pal 1993]. It can also be viewed as a preprocessing step for object detection/recognition. It is not very surprising to see that even in computer vision earlier attempts were drawn towards single aspects of the segmentation like edge detection [Marr 1980, Canny 1986, Lindeberg 1998] or grouping homogeneous regions by clustering [Coleman 1979]. The performance limitations of both these approaches independently have led to the emergence of solutions that reconsidered at the problem as a juxtaposition of both edge detection and homogeneous region grouping with implicit consideration for scale. The review paper by [Freixenet 2002] presents various approaches that attempted in merging edge based information and clustering based information in a sequential or parallel manner. The state of the art techniques that are successful in formulating the combined approach are variants of graph cuts [Shi 2000], active contours, and level sets. At the bottom of all such approaches is the definition of an optimisation scheme that seeks to find a solution under constraints such as, e.g., smoothness or minimising a measure of total energy. These approaches are much better in terms of meeting human defined ground truth compared to simpler variants involving discontinuity detection or clustering alone. The performance of computer vision approaches to image partitioning has been boosted recently by numerous contributions utilizing DCNNs for segmentation (e.g., [Noh 2015, Hong 2015, Hong 2016]). The basic structure of the encoder component of segmentation networks is similar to the hierarchical networks trained for object recognition [Krizhevsky 2012]. For example, the *AlexNet* has been trained by learning a hierarchy of kernels in the convolutional layers to extract rich feature sets for recognition from a large database of object classes. Segmentation networks [Noh 2015, Hong 2015] have been designed by adding a decoder scheme to expand the activations in the category layers through a sequence of deconvolutions steps such as in autoencoder networks [Hinton 2006a]. Even more extended versions include a mechanism of focused attention to more selectively guide the training process using class labels or segmentations [Hong 2016]. The hierarchical structure of such approaches shares several features of cortical processing through a sequence of areas with cells that increase their response selectivity at the size of their receptive fields over different stages in the cortical hierarchy. However, the explicit unfolding of the data representation in the deconvolution step to upscale to full image resolution, the specific indexing of pixel locations to invert the pooling in the deconvolution, and the large amount of training data are not biologically plausible.

A major challenge is still how to compare the validity and the quality of segmentation approaches. Recent attempts emphasise to compare the computational results - from operations on different scales - with the results of hand-drawn segmentations by human subjects [Fowlkes 2007, Arbelaez 2011]. These approaches suggest possible measures in judging the quality of automatic segmentation given that ground truth data is missing. However, the human segmentation data does not elucidate

the mechanisms underlying the processes to arrive at such partitions. Instead of a global partitioning of the visual scene, the visual system seems to adopt different strategies of computation to arrive at a meaningful segmentation of figural items. The grouping of elements into coherent form is instantiated by selectively enhancing the activity of neurons that represent the target region via a modulatory input from higher cortical stages [Lamme 1995, Lamme 1998]. The notion of feedback to contribute in the segmentation of visual scenes has been elucidated above. Recent computer vision algorithms begin to make use of such recurrent mechanisms as well. For example, since bottom-up data-driven segmentation is usually incomplete and ambiguous the use of higher-level representations might help to validate initial instances and further stabilise their representation [Ullman 2007, Borenstein 2008]. Along this line, top-down signalling applies previously acquired information about object shape (e.g., through learning), making use of the discriminative power of fragments of intermediate size, and combines this information with a hierarchy of initial segments [Ullman 2002]. Combined contour and region processing mechanisms have also been suggested to guide the segmentation. In [Arbelaez 2011], multi-scale boundaries are extracted which later prune the contours in a watershed region-filling algorithm. Algorithms of figure-ground segregation and border-ownership computation have been developed for computer vision applications to operate under realistic imaging conditions [Stein 2009, Sundberg 2011]. These were designed to solve tasks like shape detection against structured background and for video editing. Still, the robust segmentation of an image into corresponding surface patches is hard to accomplish in a reliable fashion. Performance of such methods mentioned above depends on parametrization and the unknown complexity and properties of the viewed scene. Aloimonos and coworkers proposed an active vision approach that adopted biological principles like the selection and fixation on image regions that are surrounded by closed contours [Mishra 2009, Mishra 2012]. The key here is that in this approach only the fixated region (corresponding to a surface of an object or the object itself) is then segmented based on an optimization scheme using graph-cut. All image content outside the closed region contour is background w.r.t. the selected target region or object. The functionality requires an active component to relocate the gaze and a region that is surrounded by a contrast criterion in the image.

Promising bio-inspired solutions. Numerous models that account for mechanisms of contour grouping have been proposed to linking orientation selective cells [Grossberg 1985, Grossberg 1997, Li 1998]. The rules of mutual support utilize a similarity metric in the space-orientation domain giving rise to a compatibility, or reliability measure [Kellman 1991] (see [Neumann 2001] for a review of generic principles and a taxonomy). Such principles migrated into computer vision approaches [Parent 1989, Medioni 2000, Kornprobst 2000] and, in turn, provided new challenges for experimental investigations [Sigman 2001, Ben-Shahar 2004]. Note that the investigation of structural connectivities in high dimensional feature spaces and their mapping onto a low-dimensional manifold lead to define a "neurogeometry" and the basic underlying mathematical principles of such structural principles [Petitot 2003, Citti 2014].

As outlined above, figure-ground segregation in biological vision segments an image or temporal sequence by boundary detection and integration followed by assigning border ownership direction and then tagging the figural component in the interior

of a circumscribed region. Evidence suggests that region segmentation by tagging the items which belong to extended regions involves feedback processing from higher stages in the cortical hierarchy [Scholte 2008]. Grossberg and colleagues proposed the FACADE theory (form-and-color-and-depth [Grossberg 1985, Grossberg 1993]) to account for a large body of experimental data, including figure-ground segregation and 3D surface perception. In a nutshell, the model architecture consists of mutually coupled subsystems, each one operating in a complementary fashion. A boundary contour system (BCS) for edge grouping is complemented by a feature contour system (FCS) which supplements edge grouping by allowing feature qualities, such as brightness, color, or depth, to spread within bounded compartments generated by the BCS.

The latter mechanism has recently been challenged by psychophysical experiments that measure subject reaction times in image-parsing tasks. The results suggest that a sequential mechanism groups, or tags, interior patches along a connected path between the fixation spot and a target probe. The speed of reaching a decision argues in favor of a spreading growth-cone mechanism that simultaneously operates over multiple spatial scales rather than the wave-like spreading of feature activities initiated from the perceptual object boundary [Jeurissen 2016]. Such a mechanism is proposed to also facilitate the assignment of figural sides to boundaries. BOwn computation has been incorporated in computer vision algorithms to segregate figure and background regions in natural images or scenes [Ren 2006, Hoiem 2011, Sundberg 2011]. Such approaches use local configurations of familiar shapes and integrate these via global probabilistic models to enforce consistency of contour and junction configurations [Ren 2006] of learning of templates from ensembles of image cues to depth and occlusion [Hoiem 2011].

Feedback mechanisms as they are discussed above, allow to build robust boundary representations such that junctions may be reinterpreted based on more global context information [Weidenbacher 2009]. The hierarchical processing of shape from curvature information in contour configurations [Rodriguez Sanchez 2012] can be combined with evidence for semi-global convex fragments or global convex configurations [Craft 2007]. Such activity is fed back to earlier stages of representation to propagate contextual evidences and quickly build robust object representations separated from the background. A first step towards combining such stage-wise processing capacities and integrating them with feedback that modulates activities in distributed representations at earlier stages of processing has been suggested in [Tschechne 2014a]. The step towards processing complex scenes from unconstrained camera images, however, still needs to be further investigated.

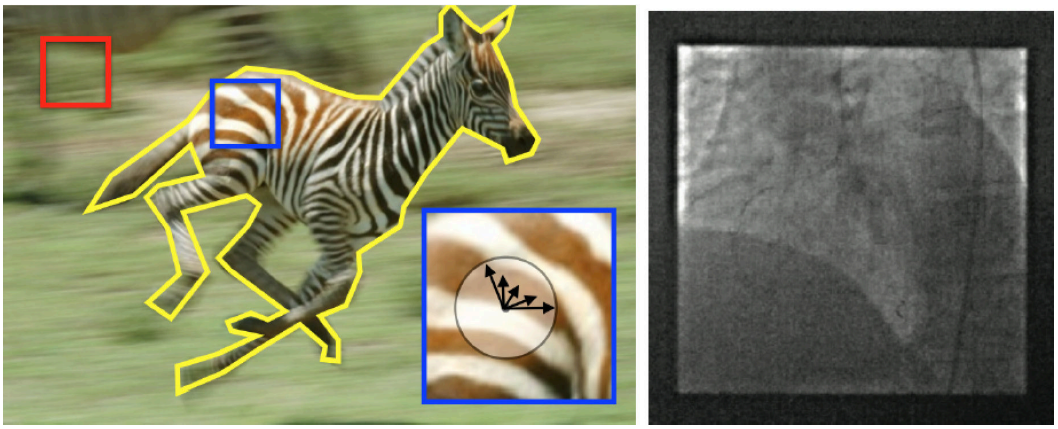
Taken together, biological vision seems to flexibly process the input in order to extract the most informative information from the optic array. The information is selected by an attention mechanism that guides the gaze to the relevant parts of the scene. It has been known for a long time that the guidance of eye movements is influenced by the observer's task of scanning pictures of natural scene content [Yarbus 1967]. More recent evidence suggests that the saccadic landing locations are guided by contraints to optimize the detection of relevant visual information from the optic array [Hayhoe 2005, Ballard 2009]. Such variability in fixation location has immediate consequences on the structure of the visual mapping into an observer representation. Consequently, segmentation might be considered as a separation problem that operates upon a high-dimensional feature space, instead of statically

separating appearances into different clusters. For example, in order to separate a target object against the background in an identification task fixation is best located approximately in the middle of the central surface region [Hayhoe 2005]. Symmetric arrangement of bounding contours (with opposite direction of BOwn) helps to select the region against the background to guide a motor action. In order to generate stable visual percept of a complex object such information must be integrated over multiple fixations [Hayhoe 1991]. In case of irregular shapes, the assignment of object belongingness requires a decision whether region elements belong to the same surface or not. Such decision-making process involves a slower sequentially operating mechanism of tracing a connecting path in a homogeneous region. Such a growth-cone mechanism has been demonstrated to act similarly on perceptual representations of contour and region representations which might tag visual elements to build a temporal signature for representations that define a connected object (compare [Jeurissen 2016]). In a different behavioral task, e.g., obstacle avoidance, the fixation close to the occluding object boundary helps to separate the optic flow pattern of the obstacle from those of the background [Raudies 2012]. Here, the obstacle is automatically selected as perceptual figure while the remaining visual scene structure and other objects more distant from the observer are treated as background. These examples demonstrate evidence that biological segmentation might be different from computer vision approaches which incorporates active selection elements building upon much more flexible and dynamic processes.

7.3.3 Optical flow

Task definition. Estimating optical flow refers to the assignment of 2-D velocity vectors at sample locations in the visual image in order to describe their displacements within the sensor's frame of reference. Such a displacement vector field constitutes the image flow representing apparent 2-D motions from their 3-D velocities being projected onto the sensor [Verri 1987, Verri 1989]. These algorithms use the change of structured light in the retinal or camera images, posing that such 2-D motions are observable from light intensity variations (and thus, are contrast dependent) due to the change in relative positions between an observer (eye or camera) and the surfaces or objects in a visual scene.

Core challenges. Achieving a robust estimation of optical flow faces several challenges. First of all, visual system has to establish form-based correspondences across temporal domain despite the fact that physical movements induced geometric and photometric distortions. Second, velocity space has to be optimally sampled and represented to achieve robust and energy efficient estimation. Third, the accuracy and reliability of the velocity estimation is dependent upon the local structure/form but the visual system must achieve a form independent velocity estimation. Difficulties arise from the fact that any local motion computation faces different sources of noise and ambiguities, such as for instance the aperture and problems. Therefore, estimating optical flow requires to resolve these local ambiguities by integrating different local motion signals while still maintaining segregated those that belong to different surfaces or objects of the visual scene (see Fig. 7.7(a)). In other words, image motion computation faces two opposite goals when computing the global object motion, integration and segmentation [Braddick 1993]. As already emphasised



(a) (b)

Figure 7.7: Core challenges in motion estimation. (a) This snapshot of a moving scenes illustrates several ideas discussed in the text: inset with the blue box shows the local ambiguity of motion estimation while the yellow boundary shows how segmentation and motion estimation are intricated. (b) One example of transparent motion encountered by computer vision, from an X-ray image (from [Auvray 2009]).

in Sec. 7.3.2, any computational machinery should be able to keep segregated the different surface/object motions since one goal of motion processing is to estimate accurately the speed and direction of each of them in order to track, capture or avoid one or several of them. Fourth, the visual system must deal with complex scenes that are full of occlusions, transparencies or non-rigid motions. This is well illustrated by the transparency case. Since optical flow is a projection of 3D displacements in the world, some situations yield to perceptual (semi-) transparency [McOwan 1996]. In videos, several causes have been identified, such as reflections, phantom special effects, dissolve effects for a gradual shot change and medical imaging such as X-rays (for example see Fig. 7.7(b)). All of these examples raise serious problems to current computer vision algorithms.

Herein, we will focus on four main computational strategies used by biological systems for dealing with the aforementioned problems. We selected them because we believe these solutions could inspire the design of better computer vision algorithms. First is *motion energy estimation* by which the visual system estimates a contrast dependent measure of translations in order to indirectly establish correspondences. Second is *local velocity estimation*: contrast dependent motion energy features must be combined to achieve a contrast invariant local velocity estimation after de-noising the dynamical inputs and resolving local ambiguities, thanks to the integration of local form and motion cues. The third challenge concerns the *global motion estimation* of each independent object, regardless its shape or appearance. Fourth, *distributed multiplexed representations* must be used by both natural and artificial systems to segment cluttered scenes, handle multiple/transparent surfaces, and encode depth ordering to achieve 3D motion perception and goal-oriented decoding.

Biological vision solution. Visual motion has been investigated in a wide range of species, from invertebrates to primates. Several computational principles have been identified as being highly conserved by evolution, as for in-

stance local motion detectors [Hassenstein 1956]. Following the seminal work of Werner Reichardt and colleagues, a huge amount of work has been achieved to elucidate the cellular mechanisms underlying local motion detection, the connectivity rules enabling optic flow detectors or basic figure-ground segmentation. Fly vision has been leading the investigation of natural image coding as well as active vision sensing. Several recent reviews can be found elsewhere (e.g. [Borst 2014, Borst 2011, Alexander 2010, Silies 2014]). In the present review, we decided to restrain the focus on the primate visual system and its dynamics. In Fig. 7.3, we have sketched the backbone of the primate cortical motion stream and its recurrent interactions with both area V1 and the 'form' stream. This figure illustrates both advantages and limits of the deep hierarchical model. Below, we will further focus on some recent data about the neuronal dynamics in regards with the four challenges identified for a better optic flow processing.

As already illustrated, the classical view of the cortical motion pathway is a feedforward cascade of cortical areas spanning from the occipital (V1) to the parietal (e.g. area VIP, area 7) lobes. This cascade forms the skeleton of the dorsal stream. Areas MT and MST are located in the deep of the superior temporal sulcus and they are considered as a pivotal hub for both object and self-motion (see, e.g., [Orban 2008, Bradley 2008, Pack 2008] for reviews). The motion pathway is extremely fast, with the information flowing in less that 20ms from the primary visual area to the frontal cortices or brainstem structures underlying visuomotor transformations (see [Lamme 2000, Bullier 2001, Masson 2012, Lisberger 2010] for reviews). These short time scales originate in the Magnocellular retino-geniculo-cortical input to area V1 carrying low spatial and high temporal frequencies luminance information with high contrast sensitivity (i.e., high contrast gain). This cortical input to layer 4β projects directly to the extra striate area MT, also called the cortical motion area. The fact that this feedforward stream by-passes the classical recurrent circuit between area V1 cortical layers is attractive for several reasons. First, it implements a fast, feedforward hierarchy fitting the classical two-stage motion computation model [Nakayama 1985, Hildreth 1987]. Direction-selective cells in area V1 are best described as spatio-temporal filters extracting motion energy along the direction orthogonal to the luminance gradient [Emerson 1992, Conway 2003, Mante 2005]. Their outputs are integrated by MT cells to compute local motion direction and speed. Such spatio-temporal integration through the convergence of V1 inputs has three objectives: extracting motion signals embedded in noise with high precision, normalising them through centre-surround interactions and solving many of the input ambiguities such as the aperture and correspondance problems. As a consequence, speed and motion direction selectivities observed at single-cell and population levels in area MT are largely independent upon the contrast or the shape of the moving inputs [Born 2005, Bradley 2008, Orban 2008]. The next convergence stage, area MST extracts object-motion through cells with receptive fields extending up to 10 to 20 degrees (area MSTl) or optic flow patterns (e.g., visual scene rotation or expansion) that are processed with very large receptive fields covering up to 2/3 of the visual field (area MSTd). Second, the fast feedforward stream illustrates the fact that built-in, fast and highly specific modules of visual information are conserved through evolution to subserve automatic, behaviour-oriented visual processing (see, e.g. [Masson 2012, Dhande 2014, Borst 2014] for reviews). Third, this anatomical motif is a good example of a canonical circuit that implements a

sequence of basic computations such as spatio-temporal filtering, gain control and normalisation at increasing spatial scales [Rust 2006]. The final stage of all of these bio-inspired models consist in a population of neurons that are broadly selective for translation speed and direction [Simoncelli 1998, Perrone 2012] as well as for complex optical flow patterns (see e.g., [Grossberg 1999, Layton 2014] for recent examples). Such backbone can then be used to compute biological motion and action recognition [Giese 2003, Escobar 2012] similar to what was observed in human and monkey parietal cortical networks (see [Giese 2015] for a recent review).

However, recent physiological studies have shown that this feedforward cornerstone of *global motion integration* must be enriched with new properties. Figure 7.3 depitcs some of these aspects, mirroring functional connectivity and computational perspectives. First, *motion energy estimation* through a set of spatio-temporal filters was recently re-evaluated to account for the neuronal responses to complex dynamical textures and natural images. When presented with rich, naturalistic inputs, responses of both V1 complex cells and MT pattern-motion neurons become contrast invariant [Priebe 2003, Cui 2013] and more selective (i.e., their tuning is sharper) [Priebe 2003, Gharaei 2013]. Their responses become also more sparse [Vinje 2000] and more precise [Baudot 2013]. These better sensitivities could be explained by a more complex integration of inputs, through a set of adaptive, excitatory- and inhibitory-weighted filters that optimally sample the spatiotemporal frequency plane [Nishimoto 2011]. Second, centre-surround interactions are much more diverse, along many different domains (e.g. retinotopic space, orientation, direction) than originally depicted by the popular Mexican-hat model. Such diversity of centre-surround interactions in both areas V1 and MT most certainly contributes to several of the computational nonlinearities mentioned above. They involve both the classical convergence of projections from one step to the next but also the dense network of lateral interactions within V1 as well as within each extra-striate areas. These lateral interactions implement long-distance normalisation, seen as centre-surround interactions at population level [Reynaud 2012] as well as feature grouping between distant elements [Gilad 2013]. These intra- and inter-cortical areas interactions can support a second important aspect of motion integration: motion diffusion. In particular, anisotropic diffusion of local motion information can play a critical role in global motion integration by propagating reliable local motion signals within the retinotopic map [Tlapale 2010]. The exact neural implementation of these mechanisms is yet unknown but modern tools will soon allow to image, and manipulate, the dynamics of these lateral interactions. The diversity of excitatory and inhibitory inputs can explains how the aperture problem is dynamically solved by MT neurons for different types of motion inputs such as plaid patterns [Rust 2006], elongated bars or barber poles [Tsui 2010]) and they are thought to be important to encode optic flow patterns [Mineault 2012] and biological motion [Escobar 2012]. Finally, the role of feedback in this context-dependent integration of local motion has been demonstrated by experimental [Hupé 1998, Nassi 2014] and computational studies [Bayerl 2004, Bayerl 2007a] and is now addressed at the physiological level despite the considerable technical difficulties (see [Cudeiro 2006] for a review). Overall, several computational studies have shown the importance of the adaptive normalisation of spatiotemporal filters for motion perception; see [Simoncini 2012] illustrating how a generic computation (normalisation) can be adaptively tuned to match the requirement of different behaviours.

Global motion integration is only one side of the coin. As pointed out by Braddick [Braddick 1993], motion integration and segmentation works hand-in-hand to selectively group the local motion signals that belong to different surfaces. For instance, some MT neurons integrate motion signals within their receptive field only if they belong to the same contour [Huang 2007] or surface [Stoner 1992]. They can also filter out motion within the receptive field when it does not belong to the same surface [Snowden 1991, Stoner 1992], a first step for representing motion transparency or structure-from-motion in area MT [Grunewald 2002]. The fact that MT neurons can thus adaptively integrate local motion signals, and explain away others is strongly related to the fact that motion sensitive cells are most often embedded in *distributed multiplexed representations*. Indeed, most direction-selective cells are also sensitive to binocular disparity [Lappe 1996, Qian 1997, Smolyanskaya 2013], eye/head motion [Nadler 2009] and dynamical perspective cues [Kim 2015] in order to filter out motion signals from outside the plane of fixation or to disambiguate motion parallax. Thus, depth and motion processing are two intricate problems allowing the brain to compute object motion in 3D space rather than in 2D space.

Depth-motion interaction is only one example of the fact that motion pathway receives and integrates visual cues from many different processing modules [Ohshiro 2011]. This is again illustrated in Fig. 7.3, where form cues can be extracted in areas V2 and V4 and sent to area MT. Information about the spatial organisation of the scene using boundaries, colours, shapes might then be used to further refine the fast and coarse estimate of the optic flow that emerges from the V1-MT-MST backbone of the hierarchy. Such cue combination is critical to overcome classical pitfalls of the feedforward model. Noteworthy, along the hierarchical cascade, information is gathered over larger and larger receptive fields at the penalty that object boundaries and shapes are blurred. Thus, large receptive fields of MT and MST neurons can be useful for tracking large objects with the eyes, or avoiding approaching ones, but they certainly lower the spatial resolution of the estimated optic flow field. This feedforward, hierarchical processing contrasts with the sharp perception that we have of the moving scene. Mixing different spatial scales through recurrent connectivity between cortical areas is one solution [Cudeiro 2006, Gur 2015]. Constraining the diffusion of motion information along edges or within surface boundaries in certainly another as shown for texture-ground segmentation [Self 2013]. Such form-based representations play a significant role in disambiguation of motion information [Geisler 1999, McCarthy 2012, Mather 2012, Heslip 2013]. It could also play a role in setting the balance between motion integration and segmentation dynamics, as illustrated in Fig. 7.3(b).

Over the last two decades, several computational vision models have been proposed to improve optic flow estimation with a bio-inspired approach. A first step is to achieve a form-independent representation of velocity from the spatio-temporal responses from V1. A dominant computational model was proposed by Heeger and Simoncelli [Simoncelli 1998], where a linear combination of afferent inputs from V1 is followed by a non linear operation known as untuned divisive normalisation. This model, and it subsequent developments [Rust 2006, Nishimoto 2011, Simoncini 2012] replicates a variety of observations from physiology to psychophysics using simple, synthetic stimuli such as drifting grating and plaids. However, this class of models cannot resolve ambiguities in regions lacking of any 2D cues because

of the absence of diffusion mechanisms. Moreover, their normalisation and weighted integration properties are still static. These two aspects may be the reason why they do not perform well on natural movies. Feedback signals from and to MT and higher cortical areas could play a key role in reducing these ambiguities. One good example was proposed by [Bayerl 2004] where dynamical feedback modulation from MT to area V1 is used to solve the aperture problem locally. An extended model of V1-MT-MST interactions that uses centre-surround competition in velocity space was later presented by [Raudies 2011], showing good optic flow computations in the presence of transparent motion. These feedback and lateral interactions primarily play the role of context dependent diffusion operators that spread the most reliable information throughout ambiguous regions. Such diffusion mechanisms can be gated to generate anisotropic propagation, taking advantage of local form information [Tlapale 2010, Beck 2010]. An attempt at utilising these distributed representation for integrating both optic flow estimation and segmentation was proposed in [Nowlan 1994]. The same model explored the role of learning in establishing the best V1 representation of motion information, although this approach was largely ignored in optic flow models contrary to object categorisation for instance. In brief, more and more computational models of biological vision take advantages of these newly-elucidated dynamical properties to explain motion perception mechanisms. But it is not clear how these ideas perfuse to computer vision.

Comparison with computer vision solutions. The vast majority of computer vision solutions for optical flow estimation can be split into four major computational approaches (see [Sun 2010, Fortun 2015] for recent reviews). First, a constancy assumption deals with correspondence problem, assuming that brightness or color is constant across adjacent frames and assigning a cost function in case of deviation. Second, the reliability of the matching assumptions optimised using priors or a regularisation to deal with the aperture problem. Both of these solutions pose the problems as an energy function and optical flow itself is treated as an energy minimisation problem. Interestingly, a lot of recent research has been done in this area, always pushing further the limits of the state-of-the-art. This research field has put a strong emphasis on performance as a criterion to select novel approaches and sophisticated benchmarks have been developed. Since the early initiatives, current benchmarks cover a much wider variety of problems. Popular examples are the Middleburry flow evaluation [Baker 2011] and, more recently the Sintel flow evaluation [Butler 2012]. The later has important features which are not present in the Middlebury benchmark: long sequences, large motions, specular reflections, motion blur, defocus blur, and atmospheric effects.

Initial motion detection is a good example where biological and computer vision research have already converged. The correlation detector proposed by Hassenstein and Reichardt [Hassenstein 1956] serves as a reference for a velocity sensitive mechanisms to find correspondences of visual structure at image locations in consecutive temporal samples. Formal equivalence of correlation detection with a multi-stage motion energy filtering has been demonstrated [Adelson 1985]. There are now several examples of spatiotemporal filtering models that are used to extract motion energy across different scales. Initial motion detection is ambiguous since motion can locally be measured only orthogonal to an extended contrast. This is called the aperture problem and mathematically it gives an ill-posed problem to solve.

For example, in gradient-based methods, one has to estimate the two velocity components from a single equation called the optical flow constraint. In spatiotemporal energy based methods, all the spatiotemporal samples lie on a straight line in frequency space and the task is to identify a plane that passes through all of them [Bradley 2008]. Computer vision has dealt with this problem in two ways: by imposing local constraints [Lucas 1981] or by posing smoothness constrains through penalty terms [Horn 1981]. More recent approaches are attempted to fuse the two formulations [Bruhn 2005]. The penalty term plays a key role as a diffusion operator can act isotropically or anisotropically [Black 1998, Scherzer 2000, Aubert 2006]. A variety of diffusion mechanisms has been proposed so that, e.g., optical flow discontinuities could be preserved depending on velocity field variations or image structures. All these mechanisms have demonstrated powerful results regarding the successful operation in complex scenes. Computational neurosciences models also tend to rely on diffusion mechanisms too, but they differ in their formulation. A first difference stems from the fact that local motion estimation is primarily based on the spatio-temporal energy estimation. Second, the representation is distributed, allowing multiple velocities at the same location, thus dealing with layered/transparent motion. The diffusion operator is also gated based on the local form cues also relying on the uncertainty estimate which could possibly be computed using the distributed representation [Nowlan 1994].

Promising bio-inspired solutions. A modern trend in bio-inspired models of motion integration is to use more form-motion interactions for disambiguating information. This should be further exploited in computer vision models. Future research will have to integrate the growing knowledge about how diffusion processes, form-motion interaction and multiplexing of different cues are implemented and impact global motion computation [Tsui 2010, Rasch 2013, McDonald 2014]. Despite the similarities in the biological and artificial approaches to solve optical flow computation, it is important to note that there is only little interaction happening between computer vision engineers and biological vision modellers. One reason might be that biological models have not been rigorously tested on regular computer vision datasets and are therefore considered as specifically confined to laboratory conditions only. It would thus be very interesting to evaluate models such as [Simoncelli 1998, Bayerl 2007a, Brinkworth 2009, Tlapale 2011b] to identify complementary strengths and weaknesses in order to find converging lines of research investigations. Figure 7.8 illustrates work initiated in this direction where three bio-inspired models that have been tested on the Middlebury optical flow dataset [Baker 2011]. Each of these models describe a potential strategy applied by the biological visual systems to solve motion estimation problem. The first model [Solari 2015], demonstrates the applicability of a feedforward model that has been suggested for motion integration by MT neurons [Rust 2006] for estimation of optical flow by extending it into a scale-space framework and applying a linear decoding scheme for conversion of MT population activity into velocity vectors. The second model [Medathati 2015a] investigates the role of contextual adaptations depending on form based cues in feedforward pooling by MT neurons. The third model [Bouecke 2011] studies the role of modulatory feedback mechanisms in solving the aperture problem.

Some elements of the mechanisms discussed above (e.g. the early motion detec-

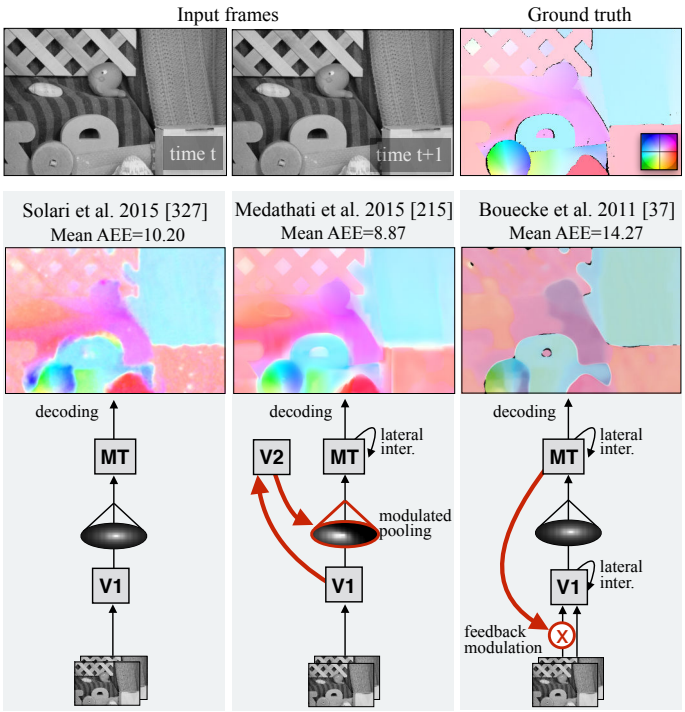


Figure 7.8: Comparison between three biological vision models tested on the Rubberwhale sequence from Middlebury dataset [Baker 2011]. First column illustrates [Solari 2015], where the authors have revisited the seminal work by Heeger and Simoncelli [Simoncelli 1998] using spatio-temporal filters to estimate optical flow from V1-MT feedforward interactions. Second column illustrates [Medathati 2015a], an extension of the Heeger and Simoncelli model with adaptive processing algorithm based on context-dependent, area V2 modulation onto the pooling of V1 inputs onto MT cells. Third column illustrates [Bouecke 2011], which incorporates modulatory feedbacks from MT to V1. Optical flow is represented using the colour-code from Middlebury dataset.

tion stage, [Heeger 1988]) have already been incorporated in recent computer vision models, For instance, the solution proposed by [Wedel 2009] uses a regularisation scheme that considers different temporal scales, namely a regular motion mechanism (using short exposure frames) as well as a slowly integrating representation (using long exposure frames), the latter resembling the form pathway in the primate visual system [Sellent 2011]. The goal there was to reduce inherent uncertainty in the input [Mac Aodha 2013]. Further constraining the computer vision models by simultaneously including some of the above-described mechanisms (e.g. tuned normalisation through lateral interactions, gated pooling to avoid estimation errors, feedback-based long range diffusion) may lead to significant improvements in optic flow processing methods and engineering solutions.

7.4 Discussion

In Sec. 7.3 we have revisited three classical computer vision tasks and discussed strategies that seemed to be used by biological vision systems in order to solve

	Reference	Model	Application	Code
Sensing	Vanrullen et al., 2002 [VanRullen 2002]	Spatial model based on difference-of-Gaussian kernels at different scales	Object recognition using the idea of latency coding	○
	Benoit et. al., 2010 [Benoit 2010]	Spatio-temporal model of retinal parvocellular and magnocellular pathways (also includes a V1 model)	Low level image processing	●
	Wohrer et al., 2009 [Wohrer 2009]	Spiking retina model with contrast gain control (*Virtual Retina*)	Comparisons to single cell recordings and large scale simulations	●
	Lorach et al., 2012, [Lorach 2012]	Retina-inspired sensor combining an asynchronous event-based light sensor (DVS) with a model pulling non-linear subunits to reproduce the parallel filtering and temporal coding of the majority of ganglion cell types	Target artificial visual systems and visual prosthetic devices	○
	Martinez et al., 2013, [Martinez-Alvarez 2013]	Compiler-based framework with an ad hoc language allowing to produce accelerated versions of the models compatible with COTS microprocessors, FPGAs or GPUs (*Retina Studio*)	Target visual prosthetic devices	○
Segmentation	Parent et al., 1989 [Parent 1989]	Model of curve detection and boundary grouping using tangent orientation and local curvature information	Tested on artificial noisy images for curve evaluation and natural images from different domains	○
	Ren et al., 2006 [Ren 2006]	Figure-ground assignment to contours in natural images based on mid-level visual shapes (so-called shapemes) and global consistency enforcement for contour junctions	Bottom-up figure-ground label assignment in still images of large data bases with human ground truth labellings	○
	Bornstein et al., 2008 [Borenstein 2008]	Model for image segmentation combining bottom-up processing (to create hierarchies of segmented uniform regions) with top-down processing (to employ shape knowledge from prior learning of image fragments)	Tested on data sets with four classes of objects to demonstrate improved segmentation and recognition performance	○
	Rodriguez et al., 2012 [Rodriguez Sanchez 2012]	Computational model of mid-level 2D shape representation utilizing hierarchical processing with end-stopping and curvature selective cells	Tested on artificial shape configurations to replicate experimental findings from neurophysiology	○
	Azzopardi et al., 2012 [Azzopardi 2012]	Computational model of center-surround and orientation selective filtering with non-linear context-dependent suppressive modulation and cross-orientation inhibition	Tested on two public data sets of natural images with contour ground truth labellings	○
	Tschechne, 2014 [Tschechne 2014a]	Recurrent network architecture for distributed multi-scale shape feature representation, boundary grouping, and border-ownership direction assignment	Tested on a selection of stimuli from public data sets	○
Optical flow	Heeger, 1988 [Heeger 1988]	Feed forward model based on spatio-temporal motion energy filters	Used to simulate psychophysical data and Yosemite sequence	○
	Nolan et al., 1994 [Nowlan 1994]	Model based on spatio-temporal motion energy filters with a selection mechanism to deal with occlusions and transparency	Optical flow estimation, tested on synthetic images only	○
	Grossberg et al., 2001 [Grossberg 2001]	Dynamical model representative of interactions between V1, V2, MT and MST areas	Grouping and optical flow estimation, tested on synthetic images only	○
	Bayerl et al., 2007 [Bayerl 2007a]	Recurrent model of V1-MT with modulatory feedbacks and a sparse coding framework for neural motion activity patterns	Optical flow estimation, tested using several real world classical videos	○
	Tlapale et al., 2010 [Tlapale 2010]	Dynamical model representative of V1-MT interactions and luminosity based motion information diffusion	Optical flow estimation, tested on synthetic images only	○
	Perrone et al., 2012 [Perrone 2012]	Model explaining the speed tuning properties of MST neurons by afferent pooling from MT	Optical flow estimation, tested on synthetic and two natural sequences	○
	Tschechne et al., 2014 [Tschechne 2014b]	Model of cortical mechanisms of motion detection using an asynchronous event-based light sensor (DVS)	Motion estimation with limited testing for action recognition	○
	Solari et al., 2015 [Solari 2015]	Multi-scale implementation of a feedforward model based on spatio-temporal motion energy filters inspired by [Heeger 1988]	Dense optical flow estimation, evaluated on Middlebury benchmark	●

Table 7.1: Prominent models for each of the three tasks considered in Sec. 7.3.

	Biological mechanism	Experimental paper	Models
Sensing	Visual adaptation	[Shapley 1984, Thoreson 2012, Kastner 2014]	[Wohrer 2009, Hérault 2010]
	Feature detection	[Kastner 2014]	[Hérault 2010]
	Sparse coding	[Pillow 2008]	[Lorach 2012]
	Precision	[Pillow 2008]	[Lorach 2012]
	Surveys	[Masland 2011, Masland 2012]	–
Segmentation	Contrast enhancement and shape representation	[Geisler 2001]	[Azzopardi 2012, Rodriguez Sanchez 2012]
	Feature integration and segmentation	[Brunswik 1953, Peterhans 1991, Field 1993, Bosking 1997, Kapadia 2000, Sigman 2001, Wolfe 2002, Li 2008, Gilad 2013]	[Grossberg 1985, Grossberg 1997, Martin 2001, Neumann 2001, Ben-Shahar 2004, Cadieu 2007, Borenstein 2008, Arbelaez 2011]
	Border ownership and figure-ground segregation	[Lamme 1995, Lamme 1998, Hupé 1998, Zhou 2000, Peterson 2008, Jeurissen 2013, Self 2013, Yang 2014]	[Grossberg 1993, Ren 2006, Craft 2007, Fowlkes 2007, Hoiem 2011, Tschechne 2014a]
	Continuation and visual routines	[Jolicoeur 1986, Kellman 1991, Poort 2012, Kogo 2013]	[Hayhoe 2005, Raudies 2010]
	Surveys	-	[Hérault 2007, Benoit 2010, Cox 2014]
Optical flow	Motion energy estimation	[Emerson 1992, Conway 2003, Mante 2005, Rust 2005]	[Adelson 1985, Heeger 1988, Simoncelli 1998]
	Local velocity estimation	[Thiele 2001b, Rust 2006, Priebe 2006, Bradley 2008, Nishimoto 2011]	[Nishimoto 2011, Solari 2015]
	Global motion integration	[Huang 2007]	[Nowlan 1994, Grossberg 2001, Bayerl 2007a, Tlapale 2010, Perrone 2012]
	Distributed multiplexed representations	[Maunsell 1983b, Basole 2003, Nadler 2009, Huk 2012, Smolyanskaya 2013]	[Buracas 1996, Lappe 1996, Qian 1997, Fernandez 2002, Ohshiro 2011]
	Surveys	[Pack 2008, Nakayama 1985]	[Bouecke 2011]

Table 7.2: Summary of the strategies highlighted in the text to solve the different task, showing where to find more details about the biological mechanisms and which models are using these strategies.

them. Tables 7.1 and 7.2 provide a concise summary of existing models for each task, together with key references about corresponding biological findings. From this meta-analysis, we have identified several research flows from biological vision that should be leveraged in order to advance computer vision algorithms. In this section, we will briefly discuss some of the major theoretical aspects and challenges described throughout the review.

7.4.1 Structural principles that relate to function

Studies in biological vision reveal structural regularities in various regions of the visual cortex. For decades, the hierarchical architecture of cortical processing has dominated, where response selectivities become more and more elaborated across levels along the hierarchy. The potential for using such deep feedforward architectures for computer vision has recently been discussed by [Kruger 2013]. However, it appears nowadays that such principles of bottom-up cascading should be combined with lateral interactions within the different cortical functional maps and the massive feedback from higher stages. We have indicated several computations (e.g., normalisation, gain control, segregation...) that could be implemented within and across functional maps by these connectivity motives. We have shown the impact of these interactions on each of the three example tasks (sensing, segmentation, optic flow) discussed throughout this article. We have also mentioned how these bio-inspired computational blocks (e.g., normalisation) can be re-used in a computer vision framework to improve image processing algorithms (e.g., statistical whitening and source separation [Lyu 2009], pattern recognition [Jarrett 2009]). One fundamental aspect of lateral and feedback interactions is that they implement context-dependent tuning of neuronal processing, over short distance (e.g. the classical centre-surround interactions) but also over much larger distances (e.g. anisotropic diffusion, feature-based attention). We have discussed the emerging ideas that these intricate, highly recurrent architectures are key ingredients to obtain an highly-flexible visual system that can be dynamically tuned to the statistics of each visual scene and to the demands of the on-going behavioural task on a moment-by-moment basis. It becomes indispensable to better understand and model how these structural principles, for which we are gaining more and more information every day, relate to functional principles. What is important in sensing, segmenting and computing optical flow is not much what could be the specific receptive fields involved in each of these problems but, rather to identify the common structural and computational architectures that they share (see Box 1). For instance, bottom-up signal representations and top-down predictions would achieve a resonant state in which the context re-enters the earlier stages of representation in order to emphasise their relevance in a larger context [Grossberg 1980, Edelman 1993]. These interactions are rooted in the generic mechanisms of response normalisation based on non-linear divisive processes. A corresponding canonical circuit, using spiking neurons representations, can then be proposed, as in [Brosch 2014] for instance. Variants of such computational elements have been used in models tackling each of these three example task; sensing, segmenting and optical flow (e.g., [Bayerl 2004, Bayerl 2007a, Wohrer 2009, Tlapale 2010]) using either functional models or neural fields formalism (see Box 1). More important, these different models can be tested on a set of real-world images and sequences taken from computer vision. This is just one exemple of the many different instances of operative solutions and algorithms that can be inspired from biology and computational vision. It is important to consider that the computational properties of a given architecture (e.g. recurrent connectivity) have been investigated in different theoretical perspectives (e.g, Kalman filtering) and different mathematical frameworks (e.g., [Rao 1999, Dimova 2009, Perrinet 2012]). Some of the biologically-plausible models assembled in Tables 7.1 offer a repertoire of realistic computational solutions

that can be a source of inspiration for novel computer vision algorithms.

7.4.2 Data encoding and representation

Biological systems are known to use several strategies such as event-based sensory processing, distributed multiplexed representation of sensory inputs and active sensory adaptation to the input statistics in order to operate in a robust and energy efficient manner. Traditionally, video inputs are captured by cameras that generate sequences of frames at a fixed rate. The consequence is that the stream of spatio-temporal scene structure is regularly sampled at fixed time steps regardless of the spatio-temporal structure. In other words, the plenoptic function [Adelson 1991] is sliced in sheets of image-like representations. The result of such a strategy is a highly redundant representation of any constant features in the scene along the temporal axis. In contrast, the brain encodes and transmits information through discrete sparse events and this spiking encoding appears at the very beginning of visual information processing, i.e., at the retina level. As discussed in Sec. 7.3.1, ganglion cells transmit a sparse asynchronous encoding of the time varying visual information to LGN and then cortical areas. This sparse event-based encoding inspired development of new type of camera sensors. Some events are registered whenever changes occur in the spatio-temporal luminance functions which are represented in a stream of events, with a location and time stamp [Lichtsteiner 2008, Liu 2010, Posch 2011]. Apart from the decrease in redundancy, the processing speed is no longer restricted to the frame-rate of the sensor. Rather, events can be delivered at a rate that is only limited by the refractory period of the sensor elements. Using these sensors brings massive improvements in terms of efficiency of scene encoding and computer vision approaches could benefit from such an alternative representation as demonstrated already on some isolated tasks.

In terms of representation, examining the richness of receptive fields of cells from retina of the visual cortex (such as in V1, MT and MST) shows that the visual system is almost always using a distributed representation for the sensory inputs. Distributed representation helps the system in a multiplicity of ways: It allows for an inherent representation for the uncertainty, it allows for task specific modulation and it could also be useful for representing the multiplicity of properties such as transparent/layered motion [Pouget 2000, Simoncelli 2001]. Another important property of biological vision that visual features are optimally encoded at the earliest stages for carrying out computations related to multiplicity of tasks in higher areas. Lastly, we have briefly mentioned that there are several codes to be used by visual networks in order to represent the complexity of natural visual scenes. Thus, it shall be very helpful to take into account this richness of representations to design systems that could deal with an ensemble of tasks simultaneously instead of subserving a single task at a time.

Recently, the application of DCNNs to solve computer vision tasks has boosted machine performance in processing complex scenes, achieving human level performance in certain scenarios. Their hierarchical structure and the utilisation of simple canonical operations (filtering, pooling, normalisation, etc.) motivated investigators to test their effectiveness in predicting cortical cell responses [Pinto 2009, Güçlü 2015]. In order to generate artificial networks with functional properties which come close to primate cortical mechanisms, a goal-diven modelling approach

has been proposed which achieved promising results [Yamins 2014]. Here, the top-layer representations should be constrained in the learning by the particular task of the whole network. The implicit assumption is that such a definition of the computational goal lies in the overlapping region of artificial and human vision systems, since otherwise the computational goals might deviate between systems as discussed above [Tsotsos 2014] (his Fig.1). The authors argue that the detailed internal structures might deviate from those identified in cortex, but additional auxiliary optimisation mechanisms might be employed to vary structures under the constraint to match the considered cortical reference system [Bergstra 2013]. The rating of any network necessitates the definition of a proper similarity measure, such as using dissimilarity measures computed from response patterns of brain regions and model representations to compare the quality of the input stimulus representations [Kriegeskorte 2009].

7.4.3 Psychophysics and human perceptual performance data

Psychophysical laws and principles which can explain large amounts of empirical observations should be further explored and exploited for designing robust vision algorithms. However, most of our knowledge about human perception has been gained using either highly artificial inputs for which the information is well-defined or natural images for which the information content is much less known. By contrast, human perception continuously adjusts information processing to the content of the images, at multiple scales and depending upon different brain states such as attention or cognition. For instance, human vision dynamically tuned decision-boundaries related to changes observed in the environment. It has been demonstrated that this adaptation can be achieved dynamically by non-linear network properties that incorporate activation transfer functions of sigmoidal shape [Grossberg 1980]. In [Chen 2010], such a principle has been adopted to define a robust image descriptor that adjusts its sensitivity to the overall signal energy, similar to human sensitivity shifts. One of the fondamental advantages of these formalism is that they can render the biological performance at many different levels, from neuronal dynamics to human performance. In other words, they can be used to adjust the algorithm parameters to different levels of constraints shared by both biological and computer vision [Tsotsos 2014]

Most of the problems in computer vision are ill-posed and observable data are insufficient in terms of variables to be estimated. In order to overcome this limitation, biological systems exploit statistical regularities. The data from human performance studies either on highly controlled stimuli with careful variations in specific attributes or large amounts of unstructured data can be used to identify the statistical regularities, particularly significant for identifying operational parameter regimes for computer vision algorithms. This strategy is already being explored in computer vision and is becoming more popular with the introduction of larges scale internet based labelling tools such as [Russell 2008, Vondrick 2013, Turpin 2014]. Classic examples for this approach in the case of scene segmentation are exploration of human marked ground truth data for static [Martin 2001] and dynamic scenes [Galasso 2013]. Thus, we advocate that further investigation on the front-end interfaces to learning functions, decision-making or separation boundaries for classifiers might improve the performance levels of existing algorithms as well as their

next generations. Emerging work such as [Scheirer 2014] illustrates the potential in this direction. [Scheirer 2014] use the human performance errors and difficulties for the task of face detection to bias the cost function of the SVM to get closer to the strategies that we might be adapting or trade-offs that our visual systems are banking on. We have provided other examples throughout the article but it is evident that further linking learning approaches with low- and mid-levels of visual information is a source of major advances in both understanding of biological vision and designing better computer vision algorithms.

7.4.4 Computational models of cortical processing

Over the last decade, many computational models have been proposed to give a formal description of phenomenological observations (e.g. perceptual decisions, population dynamics) as well as a functional description of identified circuits. Throughout this article, we have proposed that bio-inspired computer vision shall consider the existence of a few generic computational modules together with their circuit implementation. Implementing and testing these canonical operations is important to understand how efficient visual processing as well as highly flexible, task-dependent solutions can be achieved using biological circuit mechanisms and and to implement them within artificial systems. Moreover, the genericness of visual processing systems can be viewed as an emergent property from an appropriate assembly of these canonical computational blocks within a dense, highly recurrent neural networks. Computational neurosciences also investigate the nature of the representations used by these computational blocks (e.g., probabilistic population codes, population dynamics, neural maps) and we have proposed how such new theoretical ideas about neural coding can be fruitful to move forward beyond the classical isolated processing units that are typically approximated as linear-non linear filters. For each of the three example tasks, we have indicated several computational operative solutions that can be inspiring for computer vision. Table 7.1 highlights a selection of papers where even a large panels of operative solutions are described. It is beyond the scope of this paper to provide a detailed mathematical framework for each problem described or a comprehensive list of operative solutions. Still, in order to illustrate our approach, we provide in Box 1 three examples of popular operative solutions that can translate from computational to computer vision. These three examples are representative of the different mathematical frameworks described above: a functional model such as divisive normalisation that can be used for regulating population coding and decoding; a population dynamics model such as neural fields that can be used for coarse level description of lateral and feedback interactions and, lastly a neuromorphic representation data and of event-based computations such as spiking neuronal models.

The field of computational neurosciences has made enormous progress over the last decades and will be boosted by the flow of new data gathered at multiple scales, from behaviour to synapses. Testing popular computational vision models against classical benchmarks in computer vision is a first step needed to bring together these two fields of research, as illustrated above for motion processing. Translating new theoretical ideas about brain computations to artificial systems is a promising source of inspiration for computer vision as well. Both computational and computer vision share the same challenge: each one is the missing link between hardware and

behaviour, in search for generic, versatile and flexible architectures. The goal of this review was to propose some aspects of biological visual processing for which we have enough information and models to build these new architectures.

Box 1 | Three examples of operative solutions

Normalization is a generic operation present at each level of the visual processing flow, playing critical role in functions such as controlling contrast gain or tuning response selectivity [Carandini 2011]. In the context of neuronal processing, the normalization of the response R_i of a single neuron can be written by

$$R_i = \frac{{I_i}^n}{k_{tuned}{I_i}^n + \sum_j W_{ij}(I_j)^n + \sigma},$$

where $I_{\{.\}}$ indicates the net excitatory input to the neuron, $(\sum_j)$ indicates the summation over normalization pool, σ is a stabilization constant, W_{ij} are weights, n and k_{tuned} are the key parameters regulating the behavior. When $k_{tuned} = 0$ and $n = 1$ this equation represents a standard normalization. When the constant k_{tuned} is non-zero, normalization is referred to as tuned normalization. This notion has been used in computational models for, e.g., tone mapping [Meylan 2007] or optical flow [Bayerl 2004, Solari 2015].

The dynamics of biological vision results from the interaction between different cortical streams operating at different speeds but also relies on a dense network of intra-cortical and inter-cortical connections. Dynamics is generally modelled by neural fields equations which are spatially structured neural networks which represent the spatial organization of cerebral cortex [Bressloff 2012]. For example, to model the dynamics of two populations $p_1(t,r)$ and $p_2(t,r)$ (where $p.$ is the firing activity of each neural mass and r can be thought of as defining the population), a typical neural field model is

$$\frac{\partial p_1}{\partial t} = \underbrace{-\lambda_1 p_1}_{\text{decay}} + S\Big(\int_{r'} \underbrace{W_{1\to1}(t,r,r')}_{\text{lateral}} p_1(t,r') + \int_{r'} \underbrace{W_{2\to1}(t,r,r')}_{\text{feedback}} p_2(t,r') + \underbrace{K(t,r)}_{\text{external input}} \Big),$$

$$\frac{\partial p_2}{\partial t} = \underbrace{-\lambda_2 p_2}_{\text{decay}} + S\Big(\int_{r'} \underbrace{W_{1\to2}(t,r,r')}_{\text{feedforward}} p_1(t,r') + \int_{r'} \underbrace{W_{2\to2}(t,r,r')}_{\text{lateral}} p_2(t,r') \Big),$$

where the weights $W_{i\to j}$ represent the key information defining the connectivities and $S(\cdot)$ is a sigmoïdal function. Some example of neural fields model in the context of motion estimation are [Tlapale 2010, Tlapale 2011b, Rankin 2014].

Event driven processing is the basis of neural computation. A variety of equations have been proposed to model the spiking activity of single cells with different degrees of fidelity to biology [Gerstner 2002]. A simple classical case is the leaky-integrate and fire neuron (seen as a simple RC circuit) where the membrane potential u_i is given by:

$$\tau \frac{du_i}{dt} = -u_i(t) + RI_t(t),$$

with a spike emission process: the neuron i will emit a spike when $u_i(t)$ reaches a certain threshold. τ is time constant of the leaky integrator and R is the resistance of the neuron. When the neuron belongs to a network, the input current is given by $I_i(t) = \sum_j W_{j\to i} \sum_f \alpha(t - t_j^{(f)})$ where $t_j^{(f)}$ represents the time of the f-th spike of the j-th pre-synaptic neuron, $\alpha(t)$ represents the post synaptic current generated by the spike and $W_{j\to i}$ is the strength of the synaptic efficacy from neuron j to neuron i. This constitutes the building block of a spiking neural network. In term of neuromorphic architectures, this principle has inspired sensors such as event-based cameras (see Sec. 7.3.1). From a computation point of view, it has been used for biological vision [Wohrer 2009, Lorach 2012] but also for solving vision tasks [Escobar 2009, Masquelier 2010].

7.5 Conclusion

Computational models of biological vision aim at identifying and understanding the strategies used by visual systems to solve problems which are often the same as the one encountered in computer vision. As a consequence, these models would not only shed light into functioning of biological vision but also provide innovative solutions to engineering problems tackled by computer vision. In the past, these models were often limited and able to capture observations at a scale not directly relevant to solve tasks of interest for computer vision. More recently, enormous advances have been made by the two communities. Biological vision is quickly moving towards systems level understanding while computer vision has developed a great deal of task centric algorithms and datasets enabling rapid evaluation. However, computer vision engineers often ignore ideas that are not thoroughly evaluated on established datasets and modellers often limit themselves to evaluating highly selected set of stimuli. We have argued that the definition of common benchmarks will be critical to compare biological and artificial solutions as well as integrating recent advances in computational vision into new algorithms for computer vision tasks. Moreover, the identification of elementary computing blocks in biological systems and their interactions within highly recurrent networks could help resolving the conflict between task-based and generic approach of visual processing. These bio-inspired solutions could help scaling up artificial systems and improve their generalisation, their fault-tolerance and adaptability. Lastly, we have illustrated how the richness of population codes, together with some of their key properties such as sparseness, reliability and efficiency could be a fruitful source of inspiration for better representations of visual information. Overall, we argue in this review that despite their recent success, machine vision shall turn the head again towards biological vision as a source of inspiration.

Chapter 8

Conclusion

In this thesis we studied low level motion estimation in primates. We have identified that models in biological vision are heavily focussed on a subset of attributes or "readouts" related to the experiments leading to a fragmented view of the underlying processes. This introduces ambiguity in our understanding of the algorithmic and implementational strategies followed by the visual system. Intuitively, scaling up the models by taking a task centric approach could help us overcome this ambiguity.

Following this intuition, we first explored feedforward processing. We proposed an optical flow estimation algorithm by scaling up a minimal V1-MT feedforward model rooted in physiology. This involved three steps: selection of appropriate spatio-temporal frequency sampling at V1 level, appropriate velocity space sampling and decoding scheme at MT level and embedding the scheme in a scale-space framework to cover large displacements. The proposed algorithm has been benchmarked using Middlebury dataset and is publicly shared. Results demonstrated that the model could estimate optical flow on complex naturalistic stimuli but has errors at object and motion boundaries.

Motion estimation error near object and motion boundaries in the V1-MT feedforward model primarily arises due to isotropic spatial pooling of the motion energies. To minimize this kinds of errors we proposed an extension to the feedforward model by considering adaptive pooling based on the local image structure. This is done by considering inputs from V2 in the form of texture boundaries. This modification improved the performance of the model significantly and resulted in much sharper flow estimates. However, this model has limitations as the texture edge detection is noisy and linear decoding scheme that is being considered is susceptible to errors when multiple motions are present. We addressed this problem by evaluating four decoding strategies : intersection of constraints, maximum likelihood, linear regression on MT responses with in the scale-space framework and neural network based regression using multi scale-features. Considering dense sampling of the velocity space did not improve the results. However, neural network based regression has better performance hinting at the need to consider responses of the filters at different scales simultaneously.

Explorations using feedforward models gave us insights into few aspects. The V1-MT feedforward model considers pattern cells, which by their definition always combine the constituent components. Then how does the system represent transparency? Using a different sub-population of component cells? This would require that pattern and component type responses to be consistent to different kinds of stimuli. Current experimental evidence contradicts this requirement as tuning behaviour was shown switch across different kinds of stimuli such plaids/RDKs. This prompted us to explore the potential role of recurrent interactions.

We studied the role of recurrent interactions using a ring network model under neural fields formalism. By considering a structured driving input we showed

that ring network is able to reproduce prominent tuning behaviours exhibited by sub-populations of MT neurons to overlapping motion components. Compared to previous models that are rooted in feedforward weighting hypothesis, our model could explain the inherent difficulty in predicting the tuning behaviour to different input classes such as RDKs versus plaids. The integration, selection and retention behaviour exhibited by the single sub-population under recurrent interactions also challenges the assumptions made by current spatial models of integration which solve aperture problem by operating in a strong inhibition regime.

8.1 Future Work

Interfacing biological and computer vision

A rich variety of principles that have been discussed in earlier chapters could be explored further.

Psychophysics based datasets

Psychophysicists have cleverly developed stimuli that can be used to probe the algorithmic strategies adapted by the visual system. The stimuli are designed such that critical factors that could be governing algorithmic decisions could be systemically manipulated. Instead of constraining computer vision models directly using complex naturalistic scenes, it could be interesting to use carefully curated set of stimuli dealing with different aspects and progressively increasing the complexity. In the era of data driven deep learning architectures, such an approach could be useful not only to train the models but also to identify the eventual strategies adapted by the networks. Here, I elaborate this idea in case of motion stimuli.

The flow estimation can be divided into the few different challenges as illustrated in Fig.8.1

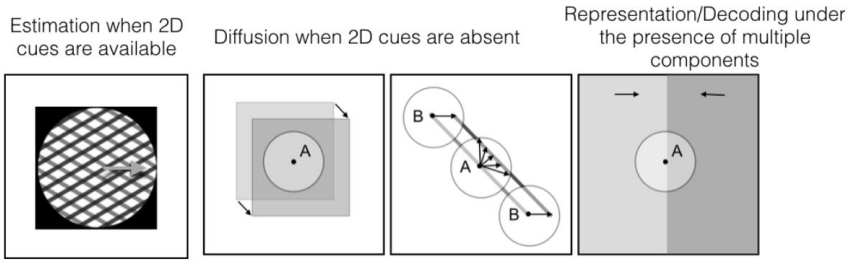


Figure 8.1: Illustrating problems faced by optical flow estimation algorithms.

- Estimation when there is no inherent ambiguity in the brightness patterns
  - How to identifying appropriate scale for analysis?
  - How to maintain contrast and illumination invariance?
- Estimation when there is no texture/2D cues available in a local region
  - How to select appropriate 2D cues?

  - How to ensure that flow field is smooth within the region of ambiguity?

- Estimation when multiple motions are involved in a local region
  - How to detection and preserve object and motion boundaries?
  - How to identify overlaid moving surfaces?

Here I give few examples of stimuli used in psychophysics that can handle the particular set of challenges described above.

- Scenario: 2D cues available without aperture effects
  - Plaid patterns generated by overlaying moving gratings and parameters such as speed of the gratings, spatial frequency of the gratings and contrast.
  - Moving Random dots patterns.
- Scenario: No texture cues present
  - Moving squares/diamonds of various sizes.
- Scenario: Locally 1D cues are present and no competing 2D cues
  - Moving bars of various lengths and orientations.
- Scenario: Locally 1D cues present but with competing 2D cues
  - Barber pole, where the line endings suggest different motions.
  - Cross barber pole.
  - Chopsticks illusion.
- Scenario: Motion or texture boundaries
  - Moving gratings with partial overlap, to check leaks in 2D cue diffusion.

Figure 8.2 illustrates some of the scenarios discussed here.

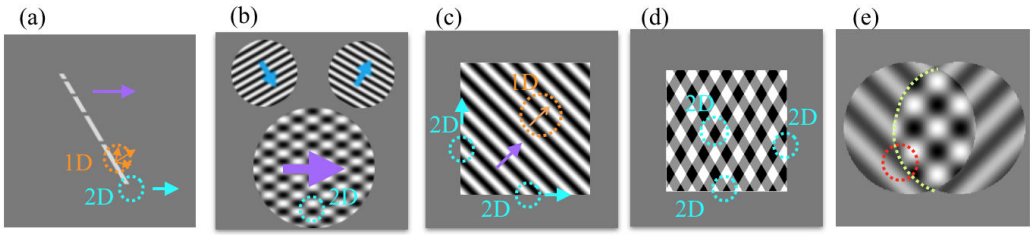


Figure 8.2: Illustrating stimuli which are representative of various scenarios visual system encounters in solving motion estimation problem. a-d: Illustration of competition between 1D and 2D cues. e: Illustration of a motion boundary.

We present the output obtained using a recent deep learning method [Dosovitskiy 2015] which has quite high performance metrics on modern computer vision datasets to a very simple stimuli such as a moving bar in Figure.8.3. We can readily observe the estimated flow is not confined to the object under motion.

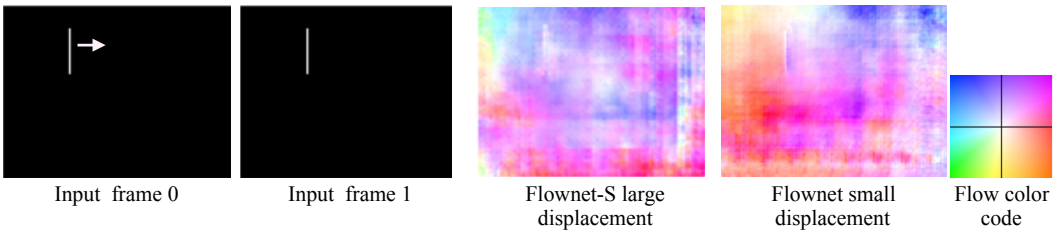


Figure 8.3: Optical flow estimation in case of a translating bar using two network configurations from [Dosovitskiy 2015].

Event driven processing

Developing camera arrays that could leverage event driven asynchronous signalling would be an interesting avenue to explore. There is a phenomenal growth in the number cameras that are present on the portable devices like cell phones, for example see Figure. 8.4. If we consider the traditional frame based cameras to construct such arrays, the processing and storage requirements are going to be very high. This can be tackled by following some of the tricks that visual system adapts such as event driven processing. For example, hybrid camera architectures could be developed following intuition from P-cells and M-cells in the retina. We could consider an array made up of high speed low resolution gray scale event driven sensors and a traditional slow speed high resolution color sensor. The complementary information coming from these different kinds of sensors could be fused to produce high speed high resolution color videos.

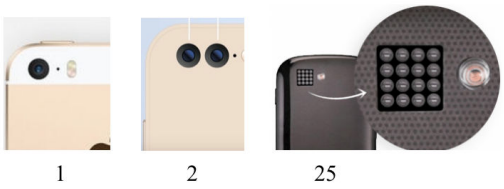


Figure 8.4: Illustrating increasing number of cameras on cell phones.

Balance in excitatory and inhibitory connections

An important aspect of the neural computation is the relative balance between excitatory and inhibitory connections. Some of the prominent regularization strategies considered in the literature can be seen in Table 8.5. Novel regularization strategies could be developed by introducing constraints such as norm of excitatory weights being equal to the norm of inhibitory weights.

Regularizer	Publication	Model
weight decay L2-norm of weights	Hinton, 1987	$\|W\|^2$ (Gaussian prior on weights)
weight eliminations generalization of weight decay	Weigand, 1991	|W|^2 / (1+|W|^2)
Sparsity L1-norm of weights	Lecun, 2006	|W| (Laplacian prior on weights)
early stopping		empirical way of Regularization in time using a statistically independent data (validation data)
DropOut	Hinton, 2012	drop activations with a bernoulli probability
DropConnect generalization of dropout	Wan, 2013	drop weights with a bernoulli probability

Figure 8.5: Regularization techniques considered in literature for training deep networks.

Neural field models

In case of the ring network model, the representation scheme developed is generic and could be extended to analyse temporal dynamics of MT neurons in other scenarios:

Increasing the number of components

In chapter 6, we have considered the case of two overlapping motion components. The study could be extended by considering additional components of motion. In the model this would mean that input is a linear combination of three gaussian bumps instead of two. Physiological data in this direction is already emerging, for example [Jazayeri 2012] have reported recording using grating triplets. Fig. 8.6 illustrates stimuli with increasing number of motion components. We made preliminary investigations into the behaviour of the ring network by considering an input with three components. The two stable solutions we found are shown in Fig. 8.7. A thorough bifurcation analysis needs to be done in order to identify all the stable solutions. It would be very interesting to see if the set of stable solution would agree with neuronal/perceptual recordings.

Temporal hyper plaids

In chapter 6, we have considered a time invariant input. Experimentally, temporal windows of motion integration have also been investigated by switching the visibility of the components along time as illustrated in Fig. 8.8. The tuning responses in such conditions could be studied by considering a periodically varying input. Early investigations of the response of the ring network with periodically switching input components reflective of the hyper plaids revealed that as the frequency of the switching increases, the network's preference to component selectivity increases. The behaviour of the network is presented in Fig. 8.9.

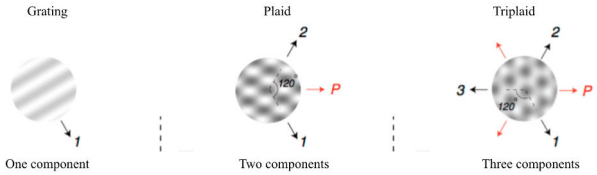


Figure 8.6: Illustrating moving grating, plaid and triplaid stimuli.

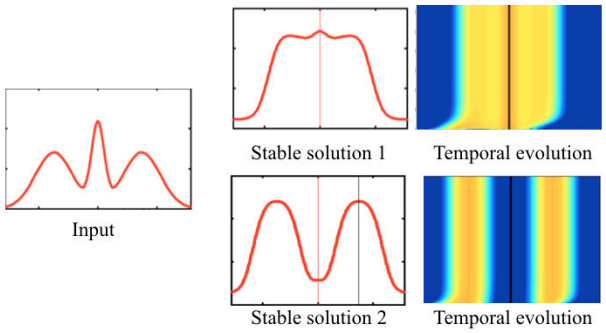


Figure 8.7: Behaviour of the ring network with input representative of three components, two stable solutions and associated temporal evolution of the population response are shown.

Developing spatialised models

The ring network model considers feature domain recurrent interactions only. Incorporation of spatial domain interactions would enable the model to deal with naturalistic videos, thus it is a natural extension to be considered. We started to work in this direction using a spatialized neural field model representative of V1 and MT populations, as shown in Fig. 8.10. Using this model we attempted to describe perceptual switching in case of barber pole stimuli. Earlier, [Rankin 2014] described such switching using feature domain representation only. Thus, their model did not provide insights into how the switching happens. Does it happen in a piecemeal way or simultaneously across the space? Using our spatialised model, we began investigating this question. So far, in our simulations, we have observed simultaneous switching across space. It needs to be further investigated to find or rule out connectivity regimes that could support switching in a piecemeal manner. Details about model could be found in [Medathati 2013]. The simulation of this model is computationally very expensive. At the continuum limit, the model requires solving around $10^6(2x256x256x64)$ unknowns. Inspired by the work by [Baladron 2013], the model has been simulated with the help of GPUs.

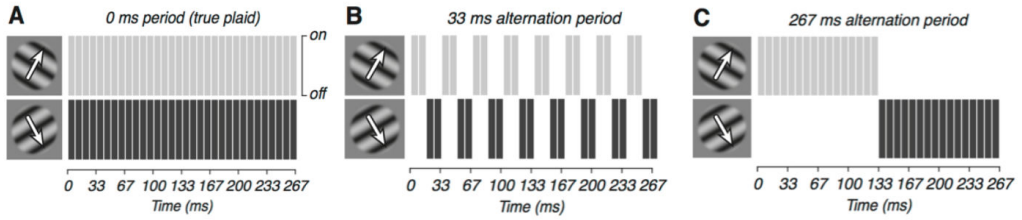


Figure 8.8: Illustrating temporal hyper-plaids, where visibility of the motion components is switched periodically. Figure adapted from [Kumbhani 2014]

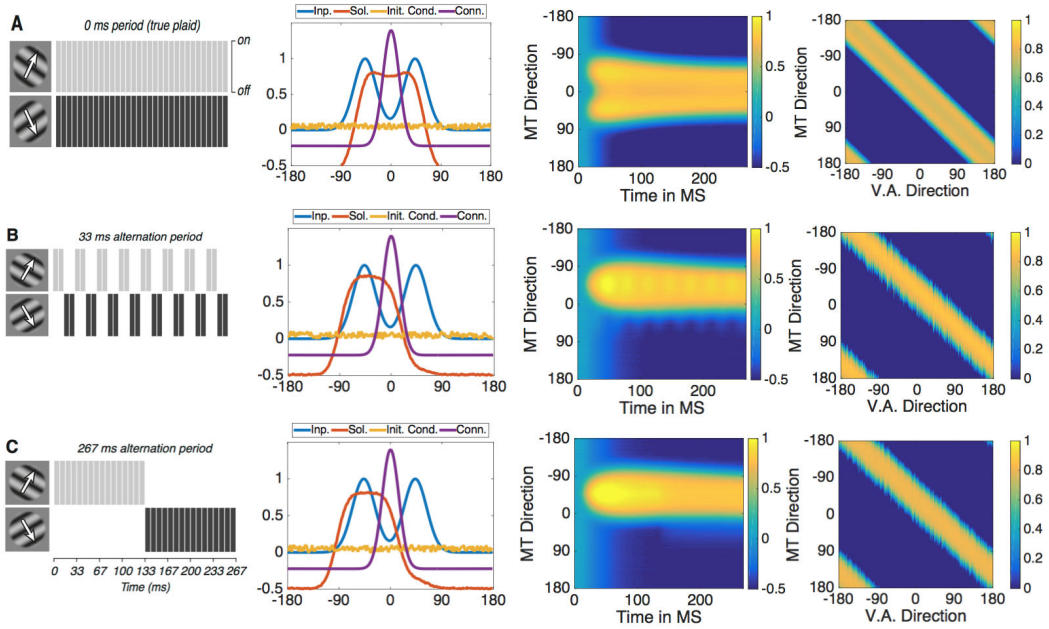


Figure 8.9: Illustrating the response selectivity of the ring network using periodically switching input representative of the hyper-plaids.

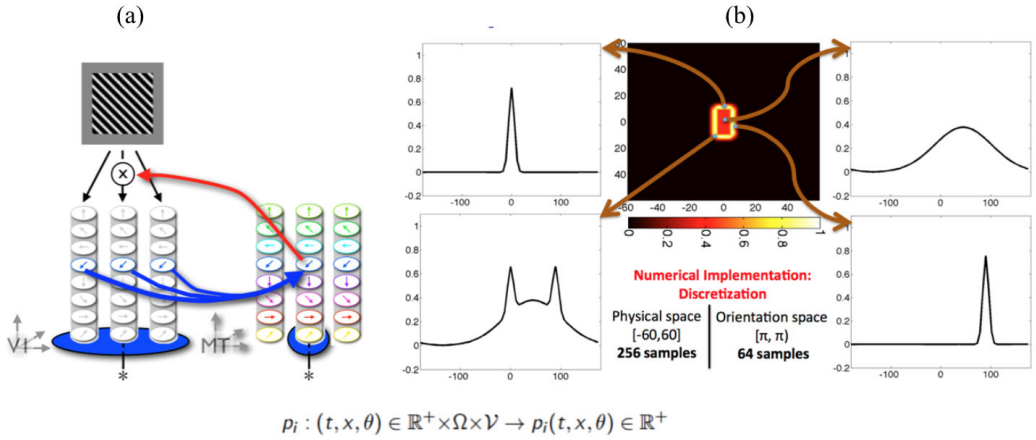


Figure 8.10: (a) A spatial neural field model representative of recurrently interacting direction selective populations of neurons from areas V1 and MT. (b) The stimuli are encoded based on the ambiguity in local motion estimation.

Publications

Refereed Journal publications

N.V. Kartheek Medathati, H. Neumann, G. S. Masson and P. Kornprobst. "**Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision**", in *Computer Vision and Image Understanding*, Volume 150, pages 1-30, 2016.

F. Solari, M. Chessa, N V Kartheek Medathati and P. Kornprobst. "**What can we expect from a classical V1-MT feedforward architecture for optical flow estimation?**", in *Image Communication*, Volume 39, Issue PB, Pages 342-354, 2015.

N. V. Kartheek Medathati, A. I. Meso, G. S. Masson, P. Kornprobst, J. Rankin. "**Understanding the impact of lateral interactions on population tuning**", *In preparation.*

Refereed International Conference proceedings

N. V. Kartheek Medathati, M. Chessa, G. S. Masson, F. Solari, P. Kornprobst. "**Decoding MT motion representation for optical flow estimation: an experimental evaluation**", *EUSIPCO*, 2015.

H. G. Cristina, N. V. Kartheek Medathati, P. Kornprobst, M. Vittorio, S. Diego, "**Improving FREAK Descriptor for Image Classification**", *ICVS*, 2015.

Conferences with abstract submission

N. V. Kartheek Medathati, A. I. Meso, G. S. Masson, P. Kornprobst, J. Rankin. "**Understanding the impact of lateral interactions on population tuning**", *AREADNE*, Santorini, Greece, 2016.

N. V. Kartheek Medathati, M. Chessa, G. S. Masson, P. Kornprobst, F. Solari. "**Adaptive Motion Pooling and Diffusion for Optical Flow**", *MODVIS workshop at VSS*, Florida, USA, 2015.

F. Solari, M. Chessa, N. V. Kartheek Medathati and P. Kornprobst. "**Benchmarking biologically inspired spatio-temporal filter based optical flow estimation on modern datasets**", *ViiHM: EPSRC network for biological and computer vision*, Stratford-Upon-Avon, UK, 2014.

C. Hilario, N. V. Kartheek Medathati, P. Kornprobst, V. Murino and D. Sona. "**A retina inspired descriptor for image classification**", *ViiHM: EPSRC network for biological and computer vision*, Stratford-Upon-Avon, UK, 2014.

N. V. Kartheek Medathati, J. Rankin, G. S. Masson and P. Kornprobst. "**Exploring the richness of center-surround dynamics: A bifurcation study**", *Bardfest: Nonlinear dynamics and stochastic methods: from neuroscience to other biological applications*, Pittsburgh, USA, 2014.

N. V. Kartheek Medathati, J. Rankin, P. Kornprobst and G. S. Masson. "**A Retinotopic neural fields model of perceptual switching in 2D motion integration**", *Bernstein Conference*, Tuebingen, Germany, 2013.

Bibliography

[Adelson 1982] E.H. Adelson and J.A. Movshon. *Phenomenal coherence of moving visual patterns*. Nature, vol. 300, no. 5892, pages 523–525, 1982. (Cited on pages 16, 30 and 35.)

[Adelson 1985] E.H. Adelson and J.R. Bergen. *Spatiotemporal energy models for the perception of motion*. Journal of the Optical Society of America A, vol. 2, pages 284–299, 1985. (Cited on pages 22, 31, 32, 44, 138 and 142.)

[Adelson 1991] Edward H. Adelson and James R. Bergen. *The Plenoptic Function and the Elements of Early Vision*. In Computational Models of Visual Processing, pages 3–20. MIT Press, 1991. (Cited on page 144.)

[Ahissar 2004] Merav Ahissar and Shaul Hochstein. *The reverse hierarchy of visual perceptual learning*. Trends in Cognitive Sciences, vol. 8, no. 10, pages 457–464, 2004. (Cited on page 109.)

[Alahi 2012] Alexandre Alahi, Raphael Ortiz and Pierre Vandergheynst. *Freak: Fast retina keypoint*. In cvpr, pages 510—517, 2012. (Cited on page 124.)

[Albright 1984] T. D. Albright. *Direction and orientation selectivity of neurons in visual area MT of the macaque*. Journal of Neurophysiology, vol. 52, no. 6, pages 1106–1030, December 1984. (Cited on page 87.)

[Albright 1987] T.D. Albright and R. Desimone. *Local precision of visuotopic organization in the middle temporal area (MT) of the macaque*. Experimental Brain Research, vol. 65, no. 3, pages 582–592, 1987. (Cited on page 54.)

[Albright 1995] Thomas D. Albright and Gene R. Stoner. *Visual Motion Perception*. Proceedings of the National Academy of Sciences of the United States of America, vol. 92, no. 7, pages 2433–2440, 1995. (Cited on page 8.)

[Alexander 2010] Borst Alexander, Haag Juergen and F. Reiff Dierk. *Fly Motion Vision*. Annual Review of Neuroscience, vol. 33, no. 1, pages 49–70, 2010. PMID: 20225934. (Cited on page 135.)

[Amari 1977] S.-I. Amari. *Dynamics of pattern formation in lateral-inhibition type neural fields*. Biological Cybernetics, vol. 27, no. 2, pages 77–87, June 1977. (Cited on page 85.)

[Andolina 2007] I.M. Andolina, H.E. Jones, W. Wang and A.M. Sillito. *Corticothalamic feedback enhances stimulus response precision in the visual system*. Proceedings of the National Academy of Sciences, vol. 104, no. 1685–1690, 2007. (Cited on page 107.)

[Andreopoulos 2013] Alexander Andreopoulos and John K. Tsotsos. *50 years of object recognition: Directions forward*. Computer Vision and Image Understanding, vol. 117, pages 827–891, 2013. (Cited on pages 113 and 119.)

[Angelucci 2003] A. Angelucci and J. Bullier. *Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons?* Journal of Physiology - Paris, vol. 97, no. 2–3, pages 141–154, 2003. (Cited on page 115.)

[Angelucci 2006] A. Angelucci and P. C. Bressloff. *Contribution of feedforward, lateral and feedback connections ot the classical receptive field center and extra-classical receptive field surround of primate V1 neurons.* Progress in Brain Research, vol. 154, pages 93–120, 2006. (Cited on page 110.)

[Anstis 1990] S. Anstis. *Imperceptible Intersections: The Chopstick Illusion.* In AI and the Eye, page 105, 1990. (Cited on page 18.)

[Anzai 2007] A. Anzai, X. Peng and D.C. Van Essen. *Neurons in monkey visual area V2 encode combinations of orientations.* Nature Neuroscience, vol. 10, no. 10, pages 1313–1321, 2007. (Cited on page 129.)

[Arbelaez 2011] Pablo Arbelaez, Michael Maire, Charless Fowlkes and Jitendra Malik. *Contour Detection and Hierarchical Image Segmentation.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pages 898–916, May 2011. (Cited on pages 128, 130, 131 and 142.)

[Arkachar 2007] Pradeep Arkachar and Meghanad D. Wagh. *Criticality of Lateral Inhibition for Edge Enhancement in Neural Systems.* Neurocomput., vol. 70, no. 4-6, pages 991–999, January 2007. (Cited on page 85.)

[Aubert 2006] Gilles Aubert and Pierre Kornprobst. Mathematical problems in image processing: partial differential equations and the calculus of variations (second edition), volume 147 of *Applied Mathematical Sciences.* Springer-Verlag, 2006. (Cited on page 139.)

[Auvray 2009] Vincent Auvray, Patrick Bouthemy and Jean Liénard. *Joint Motion Estimation and Layer Segmentation in Transparent Image Sequences—Application to Noise Reduction in X-Ray Image Sequences.* EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009. (Cited on page 134.)

[Azzopardi 2012] George Azzopardi and Nicolai Petkov. *A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model.* Biological Cybernetics, vol. 106, no. 3, pages 177–189, 2012. (Cited on pages 141 and 142.)

[Baker 2011] S Baker, D Scharstein, J P Lewis, S Roth, M J Black and R Szeliski. *A database and evaluation methodology for optical flow.* International Journal of Computer Vision, vol. 92, no. 1, pages 1–31, 2011. (Cited on pages 43, 44, 54, 56, 117, 138, 139 and 140.)

[Baladron 2013] Javier Baladron. *Exploring the neural codes using parallel hardware.* PhD thesis, Université Nice Sophia Antipolis, June 2013. (Cited on page 154.)

[Ballard 2000] D.H. Ballard, M.M. Hayhoe, G. Salgian and H. Shinoda. *Spatio-temporal organization of behavior*. Spatial Vision, vol. 13, no. 2-3, pages 321–333, 2000. (Cited on page 126.)

[Ballard 2009] D. H. Ballard and M. M. HayHoe. *Modelling the role of task in the control of gaze*. Visual Cognition, vol. 17, pages 1185–1204, 2009. (Cited on page 132.)

[Barlow 1965] H.B. Barlow and W. R. Levick. *The mechanism of directionally selective units in rabbit's retina*. J Physiol, vol. 178, pages 477–504, 1965. (Cited on page 31.)

[Barron 1994] J.L. Barron, D.J. Fleet and S.S. Beauchemin. *Performance of Optical Flow Techniques*. The International Journal of Computer Vision, vol. 12, no. 1, pages 43–77, 1994. (Cited on pages 56 and 68.)

[Basole 2003] A. Basole, L.E. White and D. Fitzpatrick. *Mapping multiple features in the population response of visual cortex*. Nature, vol. 423, pages 986–990, 2003. (Cited on page 142.)

[Bastos 2012] Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries and Karl J. Friston. *Canonical Microcircuits for Predictive Coding*. Neuron, vol. 76, no. 4, pages 695 – 711, 2012. (Cited on page 114.)

[Baudot 2013] P. Baudot, M. Levy, O. Marre, M. Pananceau and Y Fregnac. *Animation of natural scene by virtual eye-movements evoke high precision and low noise in V1 neurons*. Frontiers in Neural Circuits, vol. 7, page 206, 2013. (Cited on page 136.)

[Bayerl 2004] P. Bayerl and H. Neumann. *Disambiguating Visual Motion Through Contextual Feedback Modulation*. Neural Computation, vol. 16, no. 10, pages 2041–2066, 2004. (Cited on pages 30, 36, 38, 44, 54, 56, 68, 109, 136, 138, 143 and 147.)

[Bayerl 2007a] P. Bayerl and H. Neumann. *Disambiguating Visual Motion by Form–Motion Interaction – a Computational Model*. International Journal of Computer Vision, vol. 72, no. 1, pages 27–45, 2007. (Cited on pages 136, 139, 141, 142 and 143.)

[Bayerl 2007b] P. Bayerl and H. Neumann. *A Fast Biologically Inspired Algorithm for Recurrent Motion Estimation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 2, pages 246–260, 2007. (Cited on page 44.)

[Beck 2010] C. Beck and H. Neumann. *Interactions of motion and form in visual cortex – A neural model*. Journal of Physiology - Paris, vol. 104, pages 61–70, 2010. (Cited on pages 116 and 138.)

[Beck 2011] Cornelia Beck and Heiko Neumann. *Combining Feature Selection and Integration - A Neural Model for MT Motion Selectivity*. PLoS ONE, vol. 6, page e21254, 2011. (Cited on pages 38 and 40.)

[Ben-Shahar 2004] O. Ben-Shahar and S. Zucker. *Geometrical computations explain projection patterns of long-range horizontal connections in visual cortex*. Neural Computation, vol. 16, no. 3, pages 445—476, 2004. (Cited on pages 131 and 142.)

[Ben-Yishai 1995] R. Ben-Yishai, RL Bar-Or and H. Sompolinsky. *Theory of orientation tuning in visual cortex*. Proceedings of the National Academy of Sciences, vol. 92, no. 9, pages 3844–3848, 1995. (Cited on page 85.)

[Benoit 2009] A. Benoit, D. Alleysson, J. Hérault and P. Le Callet. *Spatio-Temporal Tone Mapping Operator based on a Retina model*. In Computational Color Imaging Workshop, 2009. (Cited on page 124.)

[Benoit 2010] A. Benoit, A. Caplier, B. Durette and J. Herault. *Using Human Visual System modeling for bio-inspired low level image processing*. Computer Vision and Image Understanding, vol. 114, no. 7, pages 758 – 773, 2010. (Cited on pages 123, 125, 141 and 142.)

[Benosman 2011] Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi and Mandyam Srinivasan. *Asynchronous frameless event-based optical flow*. Neural Networks, vol. 27, pages 32–37, 2011. (Cited on page 125.)

[Bergstra 2013] J. Bergstra, D. Yamins and D.D. Cox. *Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures*. In Proceedings of the 30th International Conference on Machine Learning (ICML), pages 115–123, Atlanta, Georgia, USA, June 2013. (Cited on pages 116 and 145.)

[Bertalmío 2014] Marcelo Bertalmío. Image processing for cinema. CRC Press, 2014. (Cited on page 124.)

[Black 1998] M.J. Black, G. Sapiro, D.H. Marimont and D. Heeger. *Robust Anisotropic Diffusion*. IEEE Trans. Imag. Proc., vol. 7, no. 3, pages 421–432, 1998. Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis. (Cited on page 139.)

[Borenstein 2008] Eran Borenstein and Shimon Ullman. *Combined Top-Down/Bottom-Up Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pages 2109–2125, 2008. (Cited on pages 126, 131, 141 and 142.)

[Born 2005] R.T. Born and D.C. Bradley. *Structure and Function of Visual Area MT*. Annu. Rev. Neurosci, vol. 28, pages 157–189, 2005. (Cited on pages 84, 111 and 135.)

[Born 2006] R.T. Born, C.C. Pack, C. Ponce and S. Yi. *Temporal Evolution of 2–Dimensional Direction Signals Used to Guide Eye Movements*. Journal of Neurophysiology, vol. 95, pages 284–300, 2006. (Cited on page 19.)

[Borst 1989] Alexander Borst and Martin Egelhaaf. *Principles of visual motion detection.* Trends in Neurosciences, vol. 12, no. 8, pages 297 – 306, 1989. (Cited on pages 30, 31 and 33.)

[Borst 2007] A. Borst. *Correlation versus gradient type motion detectors: the pros and cons.* Philosophical Transactions of the Royal Society of London: Series B, biological sciences, vol. 362, no. 1479, pages 369–374, March 2007. (Cited on page 32.)

[Borst 2011] Alexander Borst and Thomas Euler. *Seeing Things in Motion: Models, Circuits, and Mechanisms.* Neuron, vol. 71, no. 6, pages 974 – 994, 2011. (Cited on page 135.)

[Borst 2014] A. Borst. *Fly visual course control: behaviour, algorithms and circuits.* Nature Reviews Neuroscience, vol. 15, pages 590–599, 2014. (Cited on page 135.)

[Bosking 1997] W.H. Bosking, Y. Zhang, B. Schofield and D. Fitzpatrick. *Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex.* The Journal of Neuroscience, vol. 17, no. 6, pages 2112–2127, 1997. (Cited on pages 128 and 142.)

[Bouecke 2011] J.D. Bouecke, Emilien Tlapale, Pierre Kornprobst and H. Neumann. *Neural Mechanisms of Motion Detection, Integration, and Segregation: From Biology to Artificial Image Processing Systems.* EURASIP Journal on Advances in Signal Processing, vol. 2011, 2011. special issue on Biologically inspired signal processing: Analysis, algorithms, and applications. (Cited on pages 44, 116, 139, 140 and 142.)

[Braddick 1993] O. Braddick. *Segmentation versus integration in visual motion processing.* Trends in neurosciences, vol. 16, no. 7, pages 263–268, 1993. (Cited on pages 83, 115, 133 and 137.)

[Braddick 1997] O. Braddick. *Local and global representations of velocity: transparency, opponency, and global direction perception.* In Perception, volume 26, pages 995–1010. Pion Ltd., 1997. (Cited on page 84.)

[Bradley 2008] D.C. Bradley and M.S. Goyal. *Velocity computation in the primate visual system.* Nature Reviews Neuroscience, vol. 9, no. 9, pages 686–695, 2008. (Cited on pages 4, 34, 36, 44, 45, 63, 65, 76, 105, 135, 139 and 142.)

[Bressloff 2012] P.C. Bressloff. *Spatiotemporal dynamics of continuum neural fields.* Journal of Physics A: Mathematical and Theoretical, vol. 45, 2012. (Cited on page 147.)

[Briggs 2008] Farran Briggs and W. Martin Usrey. *Emerging views of corticothalamic function.* Current Opinion in Neurobiology, vol. 18, no. 4, pages 403–407, August 2008. (Cited on page 107.)

[Brinkworth 2009] Russell S. A. Brinkworth and David C. O'Carroll. *Robust Models for Optic Flow Coding in Natural Scenes Inspired by Insect Biology.*

PLoS Comput Biol, vol. 5, no. 11, page e1000555, 2009. (Cited on page 139.)

[Brosch 2014] Tobias Brosch and Heiko Neumann. *Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations*. Neural Networks, vol. 54, no. 0, pages 11 – 16, 2014. (Cited on page 143.)

[Brosch 2015a] T. Brosch, H. Neumann and P.R. Roelfsema. *Reinforcement learning of linking and tracing contours in recurrent neural networks*. PLoS Comput Biol, vol. 11, no. 10, 2015. (Cited on page 109.)

[Brosch 2015b] Tobias Brosch, Stephan Tschechne and Heiko Neumann. *On event-based optical flow detection*. Frontiers in Neuroscience, vol. 9, no. 137, April 2015. (Cited on page 125.)

[Bruhn 2005] Andrés Bruhn, Joachim Weickert and Christoph Schnörr. *Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods*. International Journal of Computer Vision, vol. 61, pages 211–231, 2005. (Cited on pages 56 and 139.)

[Brunswik 1953] E Brunswik and J Kamiya. *Ecological cue-validity of 'Proximity' and of other Gestalt factors*. The American Journal of Psychology, vol. 66, no. 1, pages 20–32, 1953. (Cited on pages 127 and 142.)

[Bullier 2001] J. Bullier. *Integrated model of visual processing*. Brain Res. Reviews, vol. 36, pages 96–107, 2001. (Cited on pages 109, 111 and 135.)

[Buracas 1996] G. T. Buracas and T. D. Albright. *Contribution of area MT to perception of three-dimensional shape: a computational study*. Vision Res, vol. 36, no. 6, pages 869–87, 1996. (Cited on page 142.)

[Buschman 2015] T.J. Buschman and S. Kastner. *From behavior to neural dynamics: an integrated theory of attention*. Neuron, vol. 88, no. 7, pages 127–144, 2015. (Cited on pages 110 and 111.)

[Butler 2012] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley and Michael J. Black. *A Naturalistic Open Source Movie for Optical Flow Evaluation*. In Proceedings of the 12th European Conference on Computer Vision - Volume Part VI, ECCV'12, pages 611–625, Berlin, Heidelberg, 2012. Springer-Verlag. (Cited on pages 43, 117 and 138.)

[Bylinskii 2015] Z. Bylinskii, E.M. DeGennaro, H. Rajalingham R.and Ruda, J. Zhang and J.K. Tsotsos. *Towards the quantitative evaluation of visual attention models*. Vision Research, vol. 116, pages 258–268, 2015. (Cited on page 110.)

[Cadieu 2007] C. Cadieu, M. Kouh, A. Pasupathy, C.E. Connor, M. Riesenhuber and T. Poggio. *A model of V4 shape selectivity and invariance*. Journal of Neurophysiology, vol. 98, no. 1733-1750, 2007. (Cited on pages 105 and 142.)

[Canny 1986] J. F. Canny. *A Computational Approach to Edge Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pages 769–798, November 1986. (Cited on page 130.)

[Carandini 1997] M. Carandini and D.L. Ringach. *Predictions of a recurrent model of orientation selectivity*. Vision Research, vol. 37, no. 21, pages 3061–3071, 1997. (Cited on page 85.)

[Carandini 2005] M. Carandini, J. B. Demb, V. Mante, D. J. Tollhurst, Y. Dan, B. A. Olshausen, J. L. Gallant and N. C. Rust. *Do we know what the early visual system does?* Journal of Neuroscience, vol. 25, no. 46, pages 10577–10597, November 2005. (Cited on page 123.)

[Carandini 2011] M. Carandini and D.J. Heeger. *Normalization as a canonical neural computation*. Nature Reviews Neuroscience, vol. 13, no. 1, pages 51–62, 2011. (Cited on pages 111, 114, 118 and 147.)

[Carandini 2012] Matteo Carandini. *From circuits to behavior: a bridge too far?* Nature Publishing Group, vol. 15, no. 4, pages 507–509, April 2012. (Cited on page 113.)

[Castet 1993] E. Castet, J. Lorenceau, M. Shiffrar and C. Bonnet. *Perceived speed of moving lines depends on orientation, length, speed and luminance*. Vision Research, vol. 33, pages 1921–1921, 1993. (Cited on page 19.)

[Cavanaugh 2002] James R. Cavanaugh, Wyeth Bair and J. Anthony Movshon. *Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons*. Journal of Neurophysiology, vol. 88, no. 5, pages 2530–2546, 2002. (Cited on page 111.)

[Chance 1999] F. S. Chance, S. B. Nelson and L. F. Abbott. *Complex cells as cortically amplified simple cells*. Nature Neuroscience, vol. 2, pages 277–282, 1999. (Cited on page 84.)

[Chaudhuri 2015] R. Chaudhuri, K. Knoblauch, M. A. Gariel, Kennedy H and X. J. Wang. *A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex*. Neuron, vol. 88, pages 419–431, 2015. (Cited on page 108.)

[Chen 2010] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen and Wen Gao. *WLD: a robust local image descriptor*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1705–1720, September 2010. (Cited on page 145.)

[Chichilnisky 2001] E. J. Chichilnisky. *A simple white noise analysis of neuronal light responses*. Network: Comput. Neural Syst., vol. 12, pages 199–213, 2001. (Cited on page 123.)

[Citti 2014] Giovanna Citti and Alessandro Sarti, editors. Neuromathematics of vision. Springer, 2014. (Cited on page 131.)

[Clausi 2000] David A. Clausi and M. Ed Jernigan. *Designing Gabor filters for optimal texture separability.* Pattern Recognition, vol. 33, no. 11, pages 1835 – 1849, 2000. (Cited on page 53.)

[Cohen 1983] M.A. Cohen and S. Grossberg. *Absolute stability of global pattern formation and parallel memory storage by competitive neural networks.* In IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, pages 815–826, 1983. (Cited on page 85.)

[Cohen 1990] Michael A. Cohen. *The stability of sustained oscillations in symmetric cooperative-competitive networks.* Neural Networks, vol. 3, no. 6, pages 609 – 612, 1990. (Cited on page 85.)

[Cohen 2009] M.R. Cohen and J.H. Maunsell. *Attention improves performance primarily by reducing interneuronal correlations.* Nature Neuroscience, vol. 12, pages 1594–1600, 2009. (Cited on page 110.)

[Coleman 1979] G.B. Coleman and Harry C. Andrews. *Image segmentation by clustering.* Proceedings of the IEEE, vol. 67, no. 5, pages 773–785, 1979. (Cited on page 130.)

[Conway 2003] B. Conway and M. Livingstone. *Space-Time Maps and Two-Bar Interactions of Different Classes of Direction-Selective Cells in Macaque V1.* Journal of Neurophysiology, vol. 89, pages 2726–2742, 2003. (Cited on pages 135 and 142.)

[Coombes 2005] Stephen Coombes. *Waves, bumps, and patterns in neural fields theories.* Biological Cybernetics, vol. 93, no. 2, pages 91–108, 2005. (Cited on page 85.)

[Coultrip 1992] Robert Coultrip, Richard Granger and Gary Lynch. *A Cortical Model of Winner-take-all Competition via Lateral Inhibition.* Neural Netw., vol. 5, no. 1, pages 47–54, January 1992. (Cited on page 85.)

[Cox 2014] David Daniel Cox and Thomas Dean. *Neural Networks and Neuroscience-Inspired Computer Vision.* Current Biology, vol. 24, no. 18, pages 921–929, 2014. (Cited on pages 103, 105, 106, 118 and 142.)

[Craft 2007] E. Craft, H. Schutze, E. Niebur and R. von der Heydt. *A neural model of figure-ground organization.* Journal of Neurophysiology, vol. 97, pages 4310–4326, 2007. (Cited on pages 132 and 142.)

[Cristobal 2015] Gabriel Cristobal, Laurent Perrinet and Matthias .S Keil, editors. Biologically inspired computer vision: Fundamentals and applications. Wiley-VCH, 2015. (Cited on pages 103 and 113.)

[Cruz-Martin 2014] A. Cruz-Martin, R.N. El-Danaf, F. Osakada, B. Sriram, O.S. Dhande, P.L. Nguyen, E.M. Callaway, A. Ghosh and A.D. Huberman. *A dedicated circuit links direction-selective retinal ganglion cells to the primary visual cortex.* Nature, vol. 507, pages 358–361, 2014. (Cited on page 114.)

[Cudeiro 2006] Javier Cudeiro and Adam M. Sillito. *Looking back: corticothalamic feedback and early visual processing.* Trends in Neurosciences, vol. 29, no. 6, pages 298–306, June 2006. (Cited on pages 107, 110, 136 and 137.)

[Cui 2013] Y. Cui, L.D. Liu, F.A. Khawaja, C.C. Pack and D.A. Butts. *Diverse suppressive influences in area MT and selectivity to complex motion features.* Journal of Neuroscience, vol. 33, no. 42, pages 16715–16728, 2013. (Cited on page 136.)

[Curtu 2004] R. Curtu and B. Ermentrout. *Pattern Formation in a Network of Excitatory and Inhibitory Cells with Adaptation.* SIAM Journal on Applied Dynamical Systems, vol. 3, page 191, 2004. (Cited on page 87.)

[D. Marr 1981] S. Ullman D. Marr. *Directional Selectivity and its Use in Early Visual Processing.* Proceedings of the Royal Society of London. Series B, Biological Sciences, vol. 211, no. 1183, pages 151–180, 1981. (Cited on pages 31 and 33.)

[Daugman 1985] John G. Daugman. *Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters.* J. Opt. Soc. Am. A, vol. 2, no. 7, pages 1160–1169, 1985. (Cited on page 44.)

[Daugman 1988] J.G. Daugman. *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 36, no. 7, pages 1169–1179, 1988. (Cited on page 112.)

[Dayan 2001] P. Dayan and L.F. Abbott. Theoretical neuroscience : Computational and mathematical modeling of neural systems. MIT Press, 2001. (Cited on page 6.)

[DeAngelis 1995] G.C. DeAngelis, I. Ohzawa and R.D. Freeman. *Receptive-field dynamics in the central visual pathways.* Trends in Neurosciences, vol. 18, no. 10, pages 451–458, 1995. (Cited on pages 44, 45 and 53.)

[Demb 2008] Jonathan B. Demb. *Functional circuitry of visual adaptation in the retina.* The Journal of Physiology, vol. 586, no. 18, pages 4377–4384, 2008. (Cited on page 121.)

[DeYoe 1988] E.A. DeYoe and D. C. Van Essen. *Concurrent processing streams in monkey visual cortex.* Trends in Neurosciences, vol. 11, no. 219-226, 1988. (Cited on pages 104 and 108.)

[Dhande 2014] O.S. Dhande and A.D. Huberman. *Retinal ganglion cell maps in the brain: implications for visual processing.* Current Opinion in Neurobiology, vol. 24, pages 133–142, 2014. (Cited on page 135.)

[Dimova 2009] Kameliya Dimova and Michael Denham. *A neurally plausible model of the dynamics of motion integration in smooth eye pursuit based on recursive Bayesian estimation.* Biological Cybernetics, vol. 100, no. 3, pages 185–201, 2009. (Cited on pages 114 and 143.)

[Dosovitskiy 2015] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazrba, V. Golkov, P. v.d. Smagt, D. Cremers and T. Brox. *FlowNet: Learning Optical Flow with Convolutional Networks*. In IEEE International Conference on Computer Vision (ICCV), 2015. (Cited on pages 151 and 152.)

[Eckhorn 1990] R Eckhorn, H Reitboeck, M Arndt and P Dicke. *Feature Linking via Synchronization among Distributed Assemblies: Simulations of Results from Cat Visual Cortex*. Neural Computation, vol. 2, no. 3, pages 293–307, 1990. (Cited on page 111.)

[Edelman 1993] Gerald M. Edelman. *Neural Darwinism: Selection and reentrant signaling in higher brain function*. Neuron, vol. 10, no. 2, pages 115 – 125, 1993. (Cited on pages 128 and 143.)

[Eilertsen 2013] G. Eilertsen, R. Wanat, R.K. Mantiuk and J. Unger. *Evaluation of Tone Mapping Operators for HDR-Video*. Computer Graphics Forum, vol. 32, no. 7, pages 275–284, October 2013. (Cited on page 124.)

[Ellias 1975] SamuelA. Ellias and Stephen Grossberg. *Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks*. Biological Cybernetics, vol. 20, no. 2, pages 69–98, 1975. (Cited on pages 85 and 86.)

[Emerson 1992] R.C. Emerson, J.R. Bergen and E.H. Adelson. *Directionally selective complex cells and the computation of motion energy in cat visual cortex*. Vision Research, vol. 32, pages 203–218, 1992. (Cited on pages 135 and 142.)

[Engel 2001] Andreas K. Engel and Wolf Singer. *Temporal binding and the neural correlates of sensory awareness*. Trends in Cognitive Sciences, vol. 5, no. 1, pages 16 – 25, 2001. (Cited on page 111.)

[Enroth-Cugell 1984] C. Enroth-Cugell and J.G. Robson. *Functional characteristics and diversity of cat retinal ganglion cells. Basic characteristics and quantitative description*. Investigative Ophthalmology and Visual Science, vol. 25, no. 250-257, 1984. (Cited on page 105.)

[Ermentrout 1992] Bard Ermentrout. *Complex Dynamics in Winner-take-all Neural Nets with Slow Inhibition*. Neural Netw., vol. 5, no. 3, pages 415–431, March 1992. (Cited on page 85.)

[Escobar 2009] Maria-Jose Escobar, Guillaume S. Masson, Thierry Viéville and Pierre Kornprobst. *Action Recognition Using a Bio-Inspired Feedforward Spiking Network*. International Journal of Computer Vision, vol. 82, no. 3, page 284, 2009. (Cited on page 147.)

[Escobar 2012] Maria-Jose Escobar and Pierre Kornprobst. *Action recognition via bio-inspired features: The richness of center-surround interaction*. Computer Vision and Image Understanding, vol. 116, no. 5, pages 593—605, 2012. (Cited on page 136.)

[Fairhall 2014] A. Fairhall. *The receptive field is dead. Long life the receptive field?* Current Opinion in Neurobiology, vol. 25, pages 9–12, 2014. (Cited on page 64.)

[Felleman 1991] D.J. Felleman and D.C. Van Essen. *Distributed hierarchical processing in the primate cerebral cortex.* Cereb Cortex, vol. 1, pages 1–47, 1991. (Cited on page 5.)

[Fennema 1979] C.L. Fennema and W.B. Thompson. *Velocity determination in scenes containing several moving objects.* Computer Graphics and Image Processing, vol. 9, no. 4, pages 301–315, 1979. (Cited on pages 4, 6 and 33.)

[Fernandez 2002] J.M. Fernandez, B. Watson and N. Qian. *Computing relief structure from motion with a distributed velocity and disparity representation.* Vision Research, vol. 42, no. 7, pages 863–898, 2002. (Cited on page 142.)

[Ferradans 2011] S. Ferradans, M. Bertalmio, E. Provenzi and V. Caselles. *An Analysis of Visual Adaptation and Contrast Perception for Tone Mapping.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 10, pages 2002–2012, October 2011. (Cited on page 124.)

[Field 1993] D.J. Field, A. Hayes and R.F. Hess. *Contour integration by the human visual system: evidence for a local "association field".* Vision Research, vol. 33, no. 2, pages 173–193, 1993. (Cited on pages 128 and 142.)

[Field 2007] G.D. Field and E.J. Chichilnisky. *Information processing in the primate retina: circuitry and coding.* Annual Review of Neuroscience, vol. 30, pages 1–30, 2007. (Cited on page 106.)

[Fortun 2015] D. Fortun, P. Bouthemy and C. Kervrann. *Optical flow modeling and computation: a survey.* Computer Vision and Image Understanding, vol. 134, pages 1–21, 2015. (Cited on pages 43 and 138.)

[Fowlkes 2007] C.C. Fowlkes, D.R. Martin and J. Malik. *Local figure-ground cues are valid for natural images.* J. of Vision, vol. 7, no. 8, pages 1–9, 2007. (Cited on pages 130 and 142.)

[Freeman 2013] Jeremy Freeman, Cosey M Ziemba, David J Heeger, Eero P Simoncelli and J Anthony Movshon. *A functional and perceptual signature of the second visual area in primates.* Nature Neuroscience, vol. 16, pages 974–981, 2013. (Cited on page 65.)

[Fregnac 2015] Y. Fregnac and B. Bathelier. *Cortical correlates of low-level perception: from neural circuits to percepts.* Neuron, vol. 88, no. 7, pages 110–126, 2015. (Cited on pages 113, 116 and 118.)

[Freixenet 2002] J. Freixenet, X. Muñoz, D. Raba, J. Martí and X. Cufí. *Yet Another Survey on Image Segmentation: Region and Boundary Information Integration.* In Anders Heyden, Gunnar Sparr, Mads Nielsen and Peter Johansen, editors, Computer Vision — ECCV 2002, volume 2352 of *Lecture*

Notes in Computer Science, pages 408–422. Springer Berlin Heidelberg, 2002. (Cited on page 130.)

[Fries 2005] P. Fries. *A mechanism for cognitive dynamics: neuronal communication through neuronal coherence*. Trends in Cognitive Science, vol. 9, pages 474–480, 2005. (Cited on pages 110 and 111.)

[Frisby 2010] John P. Frisby and James V. Stone. Seeing, second edition: The computational approach to biological vision. The MIT Press, 2nd édition, 2010. (Cited on pages 103 and 113.)

[Fukai 1997] Tomoki Fukai and Shigeru Tanaka. *A Simple Neural Network Exhibiting Selective Activation of Neuronal Ensembles: From Winner-Take-All to Winners-Share-All*. Neural Computation, vol. 9, pages 77–97, 1997. (Cited on page 85.)

[Fukushima 1987] K. Fukushima. *Neural network model for selective attention in visual pattern recognition and associative recall*. Applied Optics, vol. 26, no. 23, pages 4985–4992, 1987. (Cited on page 109.)

[Galasso 2013] F. Galasso, N.S. Nagaraja, T.J. Cardenas, T. Brox and B.Schiele. *A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis*. In IEEE International Conference on Computer Vision (ICCV), 2013. (Cited on page 145.)

[Geisler 1999] Wilson S Geisler. *Motion streaks provide a spatial code for motion direction*. Nature, vol. 400, no. 6739, pages 65–69, July 1999. (Cited on page 137.)

[Geisler 2001] W.S. Geisler, J.S. Perry, B.J. Super and D.P. Gallogly. *Edge co-occurrence in natural images predicts contour grouping performance*. Vision Research, vol. 41, pages 711–724, 2001. (Cited on pages 128 and 142.)

[Gerstner 2002] W. Gerstner and W. Kistler. Spiking neuron models. Cambridge University Press, 2002. (Cited on page 147.)

[Gharaei 2013] S. Gharaei, C. Talby, S.S. Solomon and Solomon S.G. *Texture-dependent motion signals in primate middle temporal area*. Journal of Physiology, vol. 591, pages 5671–5690, 2013. (Cited on page 136.)

[Giese 1998] Martin Giese. Dynamic neural field theory for motion perception. Springer, 1998. (Cited on page 86.)

[Giese 2003] M.A. Giese and T. Poggio. *Neural mechanisms for the recognition of biological movements and actions*. Nature Reviews Neuroscience, vol. 4, pages 179–192, 2003. (Cited on pages 108, 119 and 136.)

[Giese 2015] M.A. Giese and G. Rizzolatti. *Neural and computational mechanisms of action processing: interaction between visual and motor representations*. Neuron, vol. 88, no. 1, pages 167–180, 2015. (Cited on page 136.)

[Gilad 2013] A. Gilad, E. Meirovithz and H. Slovin. *Population responses to contour integration: early encoding of discrete elements and late perceptual grouping*. Neuron, vol. 2, no. 389–402, 2013. (Cited on pages 111, 136 and 142.)

[Gilbert 2013] Charles D Gilbert and Wu Li. *Top-down influences on visual processing*. Nature Reviews Neuroscience, vol. 14, no. 5, pages 350–363, 2013. (Cited on page 128.)

[Giuliani 2016] Massimiliano Giuliani, Xavier Lagorce, Francesco Galluppi and Ryad B. Benosman. *Event-Based Computation of Motion Flow on a Neuromorphic Analog Neural Platform*. Frontiers in Neuroscience, vol. 10, no. 9, February 2016. (Cited on page 125.)

[Gollisch 2010] T. Gollisch and M. Meister. *Eye smarter than scientists believed: neural computations in circuits of the retina*. Neuron, vol. 65, no. 2, pages 150–164, January 2010. (Cited on pages 106, 121 and 123.)

[Gorea 1991] A. Gorea and J. Lorenceau. *Directional performances with moving plaids: component-related and plaid-related processing modes coexist*. Spatial vision, vol. 5, no. 4, pages 231–252, 1991. (Cited on pages 16 and 17.)

[Granlund 1978] Goesta H. Granlund. *In search of a general picture processing operator*. Computer Graphics and Image Processing, vol. 8, no. 2, pages 155 – 173, 1978. (Cited on page 112.)

[Grossberg 1980] S. Grossberg. *How does the brain build a cognitive code?* Psychological Science, vol. 87, no. 1, pages 1–51, 1980. (Cited on pages 143 and 145.)

[Grossberg 1985] S. Grossberg and E. Mingolla. *Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading*. Psychological review, vol. 92, no. 2, pages 173–211, 1985. (Cited on pages 128, 131, 132 and 142.)

[Grossberg 1993] S. Grossberg. *A Solution of the Figure-Ground Problem for Biological Vision*. Neural Networks, vol. 6, pages 463–483, 1993. (Cited on pages 132 and 142.)

[Grossberg 1997] Stephen Grossberg, Ennio Mingolla and William D. Ross. *Visual brain and visual perception: how does the cortex do perceptual grouping?* Trends in Neurosciences, vol. 20, no. 3, pages 106–111, 1997. (Cited on pages 131 and 142.)

[Grossberg 1999] S. Grossberg, E. Mingolla and C. Pack. *A neural model of motion processing and visual navigation by cortical area MST*. Cerebral Cortex, vol. 9, no. 8, pages 878–895, December 1999. (Cited on page 136.)

[Grossberg 2001] S. Grossberg, E. Mingolla and L. Viswanathan. *Neural dynamics of motion integration and segmentation within and across apertures*. Vision Research, vol. 41, no. 19, pages 2521–2553, 2001. (Cited on pages 30, 84, 117, 141 and 142.)

[Grunewald 2002] A. Grunewald, D.C Bradley and R.A Andersen. *Neural correlates of structure-from-motion perception in macaque area V1 and MT.* Journal of Neuroscience, vol. 22, no. 14, pages 6195–6207, 2002. (Cited on page 137.)

[Grzywacz 1990] Norberto Grzywacz and A.L. Yuille. *A model for the estimate of local image velocity by cells on the visual cortex.* Proc R Soc Lond B Biol Sci., vol. 239, no. 1295, pages 129–161, March 1990. (Cited on page 44.)

[Güçlü 2015] Umut Güçlü and Marcel A. J. van Gerven. *Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream.* The Journal of Neuroscience, vol. 35, no. 27, pages 10005–10014, July 2015. (Cited on page 144.)

[Gur 2015] M. Gur. *Space reconstruction by primary visual cortex activity: a parallel, non-computational mechanism of object representation.* Trends in Neurosciences, vol. 38, no. 4, pages 207–216, 2015. (Cited on pages 109 and 137.)

[Hahnloser 1998] R.L.T. Hahnloser. *On the piecewise analysis of networks of linear threshold neurons.* Neural Networks, vol. 11, no. 4, pages 691 – 697, 1998. (Cited on page 85.)

[Hassenstein 1956] B. Hassenstein and Reichardt W. *Systemtheoretische Analyse Der Zeit, Reihenfolgen Und Vorzeichenauswertung.* In The Bewegungsperzeption Des weevil Chlorophanus. Z. Naturforsch., 1956. (Cited on pages 135 and 138.)

[Hayhoe 1991] M. Hayhoe, J. Lachter and J. Feldman. *Integration of form across saccadic eye movements.* Perception, vol. 20, pages 392–402, 1991. (Cited on page 133.)

[Hayhoe 2005] Mary Hayhoe and Dana Ballard. *Eye movements in natural behavior.* Trends in Cognitive Sciences, vol. 9, no. 4, pages 188 – 194, 2005. (Cited on pages 126, 132, 133 and 142.)

[Hedges 2011] James H Hedges, Yevgeniya Gartshteyn, Adam Kohn, Nicole C Rust, Michael N Shadlen, William T Newsome and J Anthony Movshon. *Dissociation of Neuronal and Psychophysical Responses to Local and Global Motion.* Current Biology, vol. 21, no. 23, pages 2023–2028, 2011. (Cited on page 108.)

[Heeger 1987] David J. Heeger. *Model for the extraction of image flow.* J. Opt. Soc. Am. A, vol. 4, no. 8, pages 1455–1471, 1987. (Cited on pages 44, 45 and 46.)

[Heeger 1988] D.J. Heeger. *Optical Flow Using Spatiotemporal Filters.* The International Journal of Computer Vision, vol. 1, no. 4, pages 279–302, January 1988. (Cited on pages 44, 71, 80, 140, 141 and 142.)

[Hegdé 2007a] J Hegdé and D.J. Felleman. *Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy.* The Neuroscientist, vol. 13, no. 5, pages 416–421, October 2007. (Cited on page 107.)

[Hegde 2007b] Jay Hegde and David C Van Essen. *A comparative study of shape representation in macaque visual areas V2 and V4.* Cerebral Cortex, vol. 17, no. 5, pages 1100–1116, 2007. (Cited on page 107.)

[Hérault 2007] J. Hérault and B. Durette. *Modeling Visual Perception for Image Processing.* In F. Sandoval, A. Prieto, J. Cabestany and M. Grana, editors, Computational and Ambient Intelligence : 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, 2007. (Cited on page 142.)

[Hérault 2010] J. Hérault. Vision: Images, signals and neural networks: Models of neural processing in visual perception. World Scientific, 2010. (Cited on pages 103, 113, 123, 125 and 142.)

[Heslip 2013] David Heslip, Timothy Ledgeway and Paul McGraw. *The orientation tuning of motion streak mechanisms revealed by masking.* Journal of Vision, vol. 13, no. 9, page 376, 2013. (Cited on page 137.)

[Hilario Gomez 2015] Cristina Hilario Gomez, Kartheek Medathati, Pierre Kornprobst, Vittorio Murino and Diego Sona. *Improving FREAK Descriptor for Image Classification.* In ICVS, 2015. (Cited on page 124.)

[Hildreth 1987] E C Hildreth and C Koch. *The Analysis of Visual Motion: From Computational Theory to Neuronal Mechanisms.* Annual Review of Neuroscience, vol. 10, no. 1, pages 477–533, 1987. PMID: 3551763. (Cited on pages 112 and 135.)

[Hines 2004] M. L. Hines, T. Morse, M. Migliore, N. T. Carnevale and G. M. Shepherd. *ModelDB: A Database to Support Computational Neuroscience.* J. Comput. Neurosci., vol. 17, no. 1, pages 7–11, 2004. (Cited on page 57.)

[Hinton 2006a] G.E. Hinton and R.R. Salakhutdinov. *Reducing the dimensionality of data with neural networks.* Science, vol. 313, pages 504–507, 2006. (Cited on page 130.)

[Hinton 2006b] Geoffrey E. Hinton and Simon Osindero. *A fast learning algorithm for deep belief nets.* Neural Computation, vol. 18, page 2006, 2006. (Cited on page 105.)

[Hochstein 2002] Shaul Hochstein and Merav Ahissar. *View from the Top: Hierarchies and Reverse Hierarchies in the Visual System.* Neuron, vol. 36, no. 5, pages 791 – 804, 2002. (Cited on page 109.)

[Hoiem 2011] Derek Hoiem, Alexei A. Efros and Martial Hebert. *Recovering Occlusion Boundaries from an Image.* Int. J. Comput. Vision, vol. 91, no. 3, pages 328–346, February 2011. (Cited on pages 132 and 142.)

[Hong 2015] S. Hong, H. Noh and B. Han. *Decoupled deep neural network for semi-supervised semantic segmentation.* In N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, editors, Advances in Neural Information Processing Systems (NIPS). MIT Press, 2015. (Cited on page 130.)

[Hong 2016] Seunghoon Hong, Junhyuk Oh, Bohyung Han and Honglak Lee. *Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network.* In cvpr, 2016. (Cited on page 130.)

[Horn 1981] B.K. Horn and B.G. Schunck. *Determining Optical Flow.* Artificial Intelligence, vol. 17, pages 185–203, 1981. (Cited on pages 69 and 139.)

[Huang 2007] X. Huang, T.D. Albright and G.R. Stoner. *Adaptive Surround Modulation in Cortical Area MT.* Neuron, vol. 53, pages 761–770, 2007. (Cited on pages 137 and 142.)

[Hubel 1965] D.H. Hubel and T.N. Wiesel. *Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat.* Journal of Neurophysiology, vol. 28, pages 229–289, 1965. (Cited on page 21.)

[Hubel 1968] D.H. Hubel and T.N. Wiesel. *Receptive fields and functional architecture of monkey striate cortex.* The Journal of Physiology, vol. 195, no. 1, page 215, 1968. (Cited on page 21.)

[Huk 2012] Alexander C. Huk. *Multiplexing in the primate motion pathway.* Vision Research, vol. 62, no. 0, pages 173 – 180, 2012. (Cited on page 142.)

[Hupé 1998] J.M. Hupé, A.C. James, B.R. Payne, S.G. Lomber, P. Girard and J. Bullier. *Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons.* Nature, vol. 394, pages 784–791, 1998. (Cited on pages 136 and 142.)

[Hutt 2003] A. Hutt, M. Bestehorn and T. Wennekers. *Pattern formation in intracortical neuronal fields.* Network: Computation in Neural Systems, vol. 14, pages 351–368, 2003. (Cited on page 85.)

[Ibos 2014] G. Ibos and D.J. Freedman. *Dynamic integration of task-relevant visual features in posterior parietal cortex.* Neuron, vol. 83, no. 6, pages 1468–80, 2014. (Cited on page 110.)

[Issacson 2011] J.S. Issacson and M. Scanziani. *How inhibition shapes cortical activity.* Neuron, vol. 72, pages 231–240, 2011. (Cited on page 110.)

[Itti 2001] L. Itti and C. Koch. *Computational Modelling of Visual Attention.* Nature Reviews Neuroscience, vol. 2, no. 3, pages 194–203, 2001. (Cited on page 110.)

[Jancke 2004] D. Jancke, F. Chavane, S. Naaman and A. Grinvald. *Imaging cortical correlates of illusion in early visual cortex.* Nature, vol. 428, pages 423–426, 2004. (Cited on page 111.)

[Jarrett 2009] K. Jarrett, K. Kavukcuoglu, M. Ranzato and Y. LeCun. *What is the best multi-stage architecture for object recognition?* In Computer Vision, 2009 IEEE 12th International Conference on, pages 2146–2153, 2009. (Cited on page 143.)

[Jazayeri 2012] Mehrdad Jazayeri, Pascal Wallisch and J. Anthony Movshon. *Dynamics of Macaque MT Cell Responses to Grating Triplets.* The Journal of Neuroscience, vol. 32, no. 24, pages 8242–8253, 2012. (Cited on page 153.)

[Jeurissen 2013] Danique Jeurissen, Matthew W. Self and Pieter R. Roelfsema. *Surface reconstruction, figure-ground modulation, and border-ownership.* Cognitive Neuroscience, vol. 4, no. 1, pages 50–52, 2013. PMID: 24073702. (Cited on page 142.)

[Jeurissen 2016] D. Jeurissen, M. Self and P.R. Roelfsema. *Serial grouping of 2D-image regions with object-based attention*, 2016. under revision. (Cited on pages 132 and 133.)

[Johnston 2013] Alan Johnston and Peter Scarfe. *The Role of the Harmonic Vector Average in Motion Integration.* Frontiers in Computational Neuroscience, vol. 7, no. 146, 2013. (Cited on pages 6 and 30.)

[Jolicoeur 1986] P. Jolicoeur, S. Ullman and M. Mackay. *Curve tracing: A possible basic operation the perception of spatial relations.* Memory and Cognition, vol. 14, no. 2, pages 129–140, 1986. (Cited on pages 129 and 142.)

[Jones 2012] Helen E. Jones, Ian M. Andolina, Bashir Ahmed, Stewart D. Shipp, Jake T. C. Clements, Kenneth L. Grieve, Javier Cudeiro, Thomas E. Salt and Adam M. Sillito. *Differential Feedback Modulation of Center and Surround Mechanisms in Parvocellular Cells in the Visual Thalamus.* The Journal of Neuroscience, vol. 32, no. 45, pages 15946–15951, 2012. (Cited on page 107.)

[J.R. Bergen 1984] C.H. Anderson J.R. Bergen E.H. Adelson, P.J. Burt and J.M. Ogden. *Pyramid methods in image processing.* RCA Engineer, vol. 29, pages 33–41, 1984. (Cited on page 50.)

[Kafaligonul 2015] H. Kafaligonul, B.G. Breitmeyer and H. Ogmen. *Feedforward and feedback processes in vision.* Frontiers in Psychology, vol. 6, no. 279, 2015. (Cited on page 110.)

[Kapadia 1995] M.K. Kapadia, M. Ito, C.D. Gilbert and G. Westheimer. *Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys.* Neuron, vol. 50, pages 35–41, 1995. (Cited on page 128.)

[Kapadia 2000] Mitesh K. Kapadia, Gerald Westheimer and Charles D. Gilbert. *Spatial Distribution of Contextual Interactions in Primary Visual Cortex and in Visual Perception.* Journal of Neurophysiology, vol. 84, no. 4, pages 2048–2062, 2000. (Cited on pages 128 and 142.)

[Kaski 1994] Samuel Kaski and Teuvo Kohonen. *Models of Neurodynamics and Behavior Winner-take-all networks for physiological models of competitive learning.* Neural Networks, vol. 7, no. 6, pages 973 – 984, 1994. (Cited on page 85.)

[Kastner 2014] David B Kastner and Stephen A Baccus. *Insights from the retina into the diverse and general computations of adaptation, detection, and prediction.* Current Opinion in Neurobiology, vol. 25, pages 63–69, April 2014. (Cited on pages 106, 121 and 142.)

[Kellman 1991] P.J. Kellman and T.F. Shipley. *A theory of visual interpolation in object perception.* Cognitive Psychology, vol. 23, pages 141–221, 1991. (Cited on pages 131 and 142.)

[Khorsand 2015] P. Khorsand, T. Moore and A. Soltani. *Combined contributions of feedforward and feedback inputs to bottom-up attention.* Frontiers in Psychology, vol. 6, no. 155, 2015. (Cited on page 110.)

[Kim 2014] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung and EyeWirers. *Space-time wiring specificity supports direction selectivity in the retina.* Nature, vol. 509, no. 331-336, 2014. (Cited on page 123.)

[Kim 2015] H.R. Kim, D.E. Angelaki and G.C. DeAngelis. *A novel role for visual perspective cues in the neural computation of depth.* Nature Neuroscience, vol. 18, no. 1, pages 129–137, 2015. (Cited on page 137.)

[Koenderink 1984] J.J. Koenderink. *The structure of images.* Biological Cybernetics, vol. 50, pages 363–370, 1984. (Cited on page 126.)

[Koenderink 2012] J.J. Koenderink, W. Richards and A. van Doorn. *Blow-up: a free lunch?* i-Perception, vol. 3, no. 2, pages 141–145, 2012. (Cited on page 126.)

[Koffka 1935] K. Koffka. Principles of gestalt psychology. Routledge & Kegan Paul Ltd., London, 1935. (Cited on pages 127 and 128.)

[Kogo 2013] Naoki Kogo and Johan Wagemans. *The "side" matters: How configurality is reflected in completion.* Cognitive Neuroscience, vol. 4, no. 1, pages 31–45, 2013. PMID: 24073697. (Cited on pages 127 and 142.)

[Kornprobst 2000] P. Kornprobst and G. Médioni. *Tracking segmented objects using tensor voting.* In Proceedings of the International Conference on Computer Vision and Pattern Recognition, volume 2, pages 118–125, Hilton Head Island, South Carolina, June 2000. IEEE Computer Society. (Cited on page 131.)

[Kovesi 1999] Peter Kovesi. *Image Features From Phase Congruency.* Videre: A Journal of Computer Vision Research. MIT Press, vol. 1, no. 3, 1999. (Cited on page 66.)

[Kriegeskorte 2009] N. Kriegeskorte. *Relating population-code representations between man, monkey, and computational models.* Frontiers in Neuroscience, vol. 3, no. 3, pages 363–373, 2009. (Cited on page 145.)

[Kriegeskorte 2015] Nikolaus Kriegeskorte. *Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing.* Annual Review of Vision Science, vol. 1, pages 417–446, November 2015. (Cited on page 116.)

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever and G.E. Hinton. *ImageNet classification with deep convolutional neural networks.* In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Botton and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems (NIPS), 2012. (Cited on page 130.)

[Kruger 2013] Norbert Kruger, Peter Janssen, Sinan Kalkan, Markus Lappe, Ales Leonardis, Justus Piater, Antonio J. Rodriguez-Sanchez and Laurenz Wiskott. *Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision?* IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pages 1847–1871, August 2013. (Cited on pages 103, 104, 115 and 143.)

[Kumbhani 2014] Romesh Devjibhai Kumbhani, Yasmine El-Shamayleh and J. Anthony Movshon. *Temporal and spatial limits of pattern motion sensitivity in macaque MT neurons.* Journal of Neurophysiology, 2014. (Cited on page 155.)

[Kung 2007] J. Kung, H. Yamaguchi, C. Liu, G.M. Johnson and M.D. Fairchild. *Evaluating HDR rendering algorithms.* ACM Transactions on Applied Perception, vol. 4, no. 2, July 2007. (Cited on page 124.)

[Lamme 1995] VA Lamme. *The neurophysiology of figure-ground segregation in primary visual cortex.* The Journal of Neuroscience, vol. 15, no. 2, pages 1605–1615, 1995. (Cited on pages 129, 131 and 142.)

[Lamme 1998] V.A.F. Lamme, K. Zipser and H. Spekreijse. *Figure-ground activity in primary visual cortex is suppressed by anesthesia.* PNAS, vol. 95, pages 3263–3268, 1998. (Cited on pages 131 and 142.)

[Lamme 1999] V.A.F. Lamme, V. Rodriguez-Rodriguez and H. Spekreijse. *Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey.* Cerebral Cortex, vol. 9, pages 406–413, 1999. (Cited on page 129.)

[Lamme 2000] Victor A. F. Lamme and Pieter R. Roelfsema. *The distinct modes of vision offered by feedforward and recurrent proce ssing.* Trends in Neurosciences, vol. 23, no. 11, pages 571–579, 2000. (Cited on pages 108, 110, 130 and 135.)

[Lappe 1996] Markus Lappe. *Functional consequences of an integration of motion and stereopsis in area MT of monkey extrastriate visual cortex.* Neural

Comput., vol. 8, no. 7, pages 1449–1461, 1996. (Cited on pages 137 and 142.)

[Layton 2014] O.W. Layton and N.A. Browning. *A unified model of heading and path perception in primate MSTd.* PLoS Comput Biol, vol. 10, no. 2, page e1003476, 2014. (Cited on page 136.)

[LeCun 2015] Y. LeCun, Y. Bengio and G. Hinton. *Deep learning.* Nature, vol. 521, pages 436–444, 2015. (Cited on page 118.)

[Lee 2001] T.S. Lee and M. Nguyen. *Dynamics of subjective contour formation in the early visual cortex.* Proceedings of the National Academy of Sciences, vol. 98, no. 4, page 1907, 2001. (Cited on page 107.)

[Lee 2003] T. Lee and D. Mumford. *Hierarchical Bayesian inference in the visual cortex.* J. Opt. Soc. Am. A, vol. 20, no. 7, 2003. (Cited on page 109.)

[Lehky 2013] S.R. Lehky, M.E. Sereno and A.B. Sereno. *Population coding and the labeling problem: extrinsic versus intrinsic representations.* Neural Computation, vol. 25, no. 9, pages 2235–2264, 2013. (Cited on page 112.)

[Li 1998] Z. Li. *The immersed interface method using a finite element formulation.* Applied Numerical Mathematics, vol. 27, no. 3, pages 253–267, 1998. (Cited on page 131.)

[Li 2008] W. Li, V. Piech and C.D. Gilbert. *Learning to link visual contours.* Neuron, vol. 57, pages 442–451, 2008. (Cited on pages 128 and 142.)

[Lichtsteiner 2008] Patrick Lichtsteiner, Christoph Posch and Tobi Delbruck. *A 128× 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor.* IEEE Journal of Solid-State Circuits, vol. 43, no. 2, pages 566–576, 2008. (Cited on pages 125, 126 and 144.)

[Lindeberg 1998] Tony. Lindeberg. *Feature Detection with Automatic Scale Selection.* The International Journal of Computer Vision, vol. 30, no. 2, pages 77–116, 1998. (Cited on page 130.)

[Lisberger 2010] S.G.L. Lisberger. *Visual guidance of smooth-pursuit eye movements: sensation, action and what happens in between.* Neuron, vol. 66, no. 4, pages 477–491, 2010. (Cited on page 135.)

[Liu 2009] Yuan Sophie Liu, Charles F Stevens and Tatyana O Sharpee. *Predictable irregularities in retinal receptive fields.* Proceedings of the National Academy of Sciences, vol. 106, no. 38, pages 16499–16504, September 2009. (Cited on page 125.)

[Liu 2010] S.-C. Liu and T. Delbruck. *Neuromorphic sensory systems.* Current Opinion in Neurobiology, vol. 20, pages 1–8, 2010. (Cited on pages 125 and 144.)

[Liu 2014] Dengyu Liu, Jinwei Gu, Y. Hitomi, M. Gupta, T. Mitsunaga and S.K. Nayar. *Efficient Space-Time Sampling with Pixel-Wise Coded Exposure for High-Speed Imaging*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 36, no. 2, pages 248–260, 2014. (Cited on page 123.)

[Liu 2015] Shih-Chii Liu, Tobi Delbruck, Giacomo Indiveri, Adrian Whatley and Rodney Douglas, editors. Event-based neuromorphic systems. Wiley, January 2015. (Cited on pages 103 and 125.)

[Livingstone 1988] M. Livingstone and David H Hubel. *Segregation of form, color, movement and depth: anatomy, physiology and perception*. Science, vol. 240, no. 740-749, 1988. (Cited on page 108.)

[Lorach 2012] Henri Lorach, Ryad Benosman, Olivier Marre, Sio-Hoi Ieng, José A Sahel and Serge Picaud. *Artificial retina: the multichannel processing of the mammalian retina achieved with a neuromorphic asynchronous light acquisition device*. Journal of Neural Engineering, vol. 9, no. 6, page 066004, October 2012. (Cited on pages 123, 125, 141, 142 and 147.)

[Lowe 2001] David G. Lowe. *Local feature view clustering for 3D object recognition*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 682–688, 2001. (Cited on page 125.)

[Lucas 1981] B. Lucas and T. Kanade. *An Iterative Image Registration Technique with an Application to Stereo Vision*. In International Joint Conference on Artificial Intelligence, pages 674–679, 1981. (Cited on pages 43 and 139.)

[Luo 2008] L. Luo, E.M. Callaway and K. Svoboda. *Genetic dissection of neural circuits*. Neuron, vol. 57, no. 5, pages 634–660, 2008. (Cited on page 110.)

[Lyu 2009] Siwei Lyu and Eero P. Simoncelli. *Nonlinear Extraction of Independent Components of Natural Images Using Radial Gaussianization*. Neural Comput., vol. 21, no. 6, pages 1485–1519, June 2009. (Cited on page 143.)

[Ma 2014] Wei Ji Ma and Mehrdad Jazayeri. *Neural coding of uncertainty and probability*. Annual Review of Neuroscience, vol. 37, no. 1, pages 205–220, 2014. (Cited on page 65.)

[Mac Aodha 2013] O. Mac Aodha, A. Humayun, M. Pollefeys and G.J. Brostow. *Learning a Confidence Measure for Optical Flow*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 5, pages 1107–1120, 2013. (Cited on page 140.)

[Majani 1989] E. Majani, R. Erlanson and Y. Abu-Mostafa. *On the K-winners-take-all-network*. In David S. Touretzky, editor, Advances in Neural Information Processing Systems, chapter On the K-winners-take-all-network, pages 634–642. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989. (Cited on page 85.)

[Mante 2005] Valerio Mante and Matteo Carandini. *Mapping of stimulus energy in primary visual cortex*. Journal of Neurophysiology, vol. 94, pages 788—798, March 2005. (Cited on pages 87, 135 and 142.)

[Mao 2007] Zhi-Hong Mao and S.G. Massaquoi. *Dynamics of Winner-Take-All Competition in Recurrent Neural Networks With Lateral Inhibition.* Neural Networks, IEEE Transactions on, vol. 18, no. 1, pages 55–69, 2007. (Cited on pages 85 and 86.)

[Marc 2013] R.E. Marc, B.W. Jones, C.B. Watt, J.R. Anderson, C. Sigulinsky and S. Lauritzen. *Retinal connectomics: towards complete, accurate networks.* Progress in Retinal and Eye Research, vol. 37, pages 141–162, 2013. (Cited on page 121.)

[Markov 2013] N.T. Markov, M. Ercsey-Ravasz, D.C. Van Essen, K. Knoblauch, Z. Toroczkai and H. Kennedy. *Cortical high-density counterstream architectures.* Science, vol. 342, no. 1238406-1-13, 2013. (Cited on pages 104 and 107.)

[Markov 2014] Nikola T. Markov, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, Pascal Barone, Colette Dehay, Kenneth Knoblauch and Henry Kennedy. *Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex.* Journal of Comparative Neurology, vol. 522, no. 1, pages 225–259, 2014. (Cited on page 115.)

[Marr 1980] D. Marr and E. Hildreth. *Theory of Edge Detection.* Proceedings of the Royal Society London, B, vol. 207, pages 187–217, 1980. (Cited on page 130.)

[Marr 1982] D. Marr. Vision. W.H. Freeman and Co., 1982. (Cited on pages 4, 104 and 112.)

[Marr 1983] David Marr. Vision: A computational investigation into the human representation and processing of visual information. Henry Holt & Company, June 1983. (Cited on page 4.)

[Martí 2012] D. Martí and J. Rinzel. *Dynamics of feature categorization.* Neural Computation, vol. in press, 2012. (Cited on page 85.)

[Martin 2001] D. Martin, C. Fowlkes, D. Tal and J. Malik. *A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics.* In Proc. 8th Int'l Conf. Computer Vision, volume 2, pages 416–423, 2001. (Cited on pages 126, 142 and 145.)

[Martinez-Alvarez 2013] Antonio Martinez-Alvarez, Andrés Olmedo-Payá, Sergio Cuenca-Asensi, José Manuel Ferrandez and Eduardo Fernandez. *RetinaStudio: A bioinspired framework to encode visual information.* Neurocomputing, vol. 114, pages 45–53, August 2013. (Cited on page 141.)

[Masland 2011] R H Masland. *Cell Populations of the Retina: The Proctor Lecture.* Investigative Ophthalmology and Visual Science, vol. 52, no. 7, pages 4581–4591, June 2011. (Cited on pages 121 and 142.)

[Masland 2012] Richard H. Masland. *The Neuronal Organization of the Retina.* Neuron, vol. 76, no. 2, pages 266–280, October 2012. (Cited on pages 121, 122 and 142.)

[Masquelier 2010] T. Masquelier and S.J. Thorpe. *Learning to recognize objects using waves of spikes and Spike Timing-Dependent Plasticity.* In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2010. (Cited on page 147.)

[Masson 2010] G.S. Masson and U.J. Ilg, editors. Dynamics of visual motion processing. Neuronal, Behavioral, and Computational Approaches. Springer Verlag, 1 édition, 2010. (Cited on pages 44 and 64.)

[Masson 2012] G.S. Masson and L.U. Perrinet. *The behavioral receptive field underlying motion integration for primate tracking eye movements.* Neurosciences and BioBehavioral Reviews, vol. 36, no. 1, pages 1–25, 2012. (Cited on page 135.)

[Mather 2012] George Mather, Andrea Pavan, Rosilari M. Bellacosa and Clara Casco. *Psychophysical evidence for interactions between visual motion and form processing at the level of motion integrating receptive fields.* Neuropsychologia, vol. 50, no. 1, pages 153 – 159, 2012. (Cited on pages 108 and 137.)

[Maunsell 1983a] J. H. Maunsell and D. C. Van Essen. *Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation.* Journal of Neurophysiology, vol. 49, no. 5, pages 1127–1147, 1983. (Cited on pages 48 and 75.)

[Maunsell 1983b] J.H. Maunsell and D.C. Van Essen. *Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity.* Journal of Neurophysiology, vol. 49, pages 1148–1167, 1983. (Cited on page 142.)

[McCarthy 2012] J.D. McCarthy, D. Cordeiro and G.P. Caplovitz. *Local form–motion interactions influence global form perception.* Attention, Perception, & Psychophysics, vol. 74, no. 5, pages 816–823, 2012. (Cited on page 137.)

[McDonald 2014] J. Scott McDonald, Colin W. G. Clifford, Selina S. Solomon, Spencer C. Chen and Samuel G. Solomon. *Integration and segregation of multiple motion signals by neurons in area MT of primate.* Journal of Neurophysiology, vol. 111, no. 2, pages 369–378, 2014. (Cited on pages 27, 28, 29, 84, 92, 94, 98 and 139.)

[McOwan 1996] Peter W. McOwan and Alan Johnston. *Motion transparency arises from perceptual grouping: evidence from luminance and contrast modulation motion displays.* Current Biology, vol. 6, no. 10, pages 1343 – 1346, 1996. (Cited on page 134.)

[Medathati 2013] N. V. Kartheek Medathati, James Rankin, Pierre Kornprobst and Guillaume S. Masson. *A retinotopic neural fields model of perceptual switching in 2D motion integration*. In Bernstein Conference, 2013. (Cited on page 154.)

[Medathati 2015a] N. V. Kartheek Medathati, Manuela Chessa, Guillaume S. Masson, Pierre Kornprobst and Fabio Solari. *Adaptive Motion Pooling and Diffusion for Optical Flow*. Technical report 8695, INRIA, March 2015. (Cited on pages 139 and 140.)

[Medathati 2015b] N. V. Kartheek Medathati, Heiko Neumann, Guillaume S. Masson and Pierre Kornprobst. *Bio-Inspired Computer Vision: Setting the Basis for a New Departure*. Technical report 8698, INRIA, March 2015. (Cited on page 44.)

[Medioni 2000] G. Medioni, M.S. Lee and C.K. Tang. A computational framework for segmentation and grouping. Elsevier, 2000. (Cited on page 131.)

[Merabet 1998] L. Merabet, A. Desautels, K. Minville and C. Casanova. *Motion integration in a thalamic visual nucleus*. Nature, vol. 396, no. 265–268, 1998. (Cited on page 106.)

[Meylan 2007] Laurence Meylan, David Alleysson and Sabine Süsstrunk. *Model of retinal local adaptation for the tone mapping of color filter array images*. J. Opt. Soc. Am. A, vol. 24, no. 9, pages 2807–2816, 2007. (Cited on pages 124 and 147.)

[Milner 1974] P.H. Milner. *A model for visual shape recognition*. Psychological Review, vol. 81, no. 6, pages 521–535, 1974. (Cited on page 109.)

[Milner 2008] A. D. Milner and M. A. Goodale. *Two visual systems re–viewed*. Neuropsychologia, vol. 46, pages 774–785, 2008. (Cited on page 104.)

[Mineault 2012] P.J. Mineault, F.A. Khawaja, D.A. Butts and C.C. Pack. *Hierarchical processing of complex motion along the primate dorsal visual pathways*. Proceedings of the National Academy of Sciences, vol. 109, no. 972-980, 2012. (Cited on page 136.)

[Mishra 2009] A.K. Mishra and Y. Aloimonos. *Active segmentation*. International Journal of Humanoid Robotics (IJHR), vol. 6, no. 3, pages 361—386, 2009. (Cited on page 131.)

[Mishra 2012] A.K. Mishra, Y. Aloimonos, L.-F. Cheong and A.A. Kassim. *Active visual segmentation*. pami, vol. 34, no. 4, pages 639–653, 2012. (Cited on page 131.)

[Mnih 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovsky, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg and Demis Hassabis. *Human-level control through*

deep reinforcement learning. Nature, vol. 518, pages 529–533, February 2015. (Cited on page 118.)

[Motter 1993] B. Motter. *Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli*. Journal of Neurophysiology, vol. 70, no. 3, pages 909–919, 1993. (Cited on page 110.)

[Movshon 1978] J A Movshon, I D Thompson and D J Tolhurst. *Spatial summation in the receptive fields of simple cells in the cat's striate cortex*. The Journal of Physiology, vol. 283, pages 53–77, October 1978. (Cited on page 21.)

[Movshon 1985] J.A. Movshon, E.H. Adelson, M.S. Gizzi and W.T. Newsome. *The analysis of visual moving patterns*. Pattern recognition mechanisms, pages 117–151, 1985. (Cited on pages 22, 23, 84 and 98.)

[Muchungi 2012] Kendi Muchungi and Matthew Casey. *Simulating Light Adaptation in the Retina with Rod-Cone Coupling*. In ICANN, pages 339–346. Springer Berlin Heidelberg, Berlin, Heidelberg, September 2012. (Cited on page 124.)

[Muller 2014] L. Muller, A. Reynaud, F. Chavane and A. Destexhe. *The stimulus-evoked population response in visual cortex of awake monkey is a propagating wave*. Nature Communications, vol. 5, no. 3675, 2014. (Cited on page 111.)

[Mumford 1991] D. Mumford. *On the computational architecture of the neocortex. I. The role of the thalamo-cortical loop*. Biological Cybernetics, vol. 65, pages 135–145, 1991. (Cited on page 107.)

[Nadler 2009] J.W. Nadler, M. Nawrot, D.E. Angelaki and G.C. DeAngelis. *MT neurons combine visual motion with a smooth eye movement signal to code depth-sign from motion parallax*. Neuron, vol. 63, no. 4, pages 523–532, 2009. (Cited on pages 137 and 142.)

[Nakayama 1984] Ken Nakayama. *Biological Image Motion Processing: A Review*. Vision Research, vol. 25, pages 625–660, 1984. (Cited on page 13.)

[Nakayama 1985] Ken Nakayama. *Biological image motion processing: A review*. Vision Research, vol. 25, no. 5, pages 625 – 660, 1985. (Cited on pages 135 and 142.)

[Nandy 2013] A.S. Nandy, T.O. Sharpee, J.H. Reynolds and J.F. Mitchell. *The fine structure of shape tuning in area V4*. Neuron, vol. 78, no. 1102-1115, 2013. (Cited on page 105.)

[Nasser 2013] Hassan Nasser, Selim Kraria and Bruno Cessac. *EnaS: a new software for neural population analysis in large scale spiking networks*. In Springer Series in Computational Neuroscience. Organization for Computational Neurosciences, July 2013. (Cited on page 123.)

[Nassi 2014] J.J. Nassi, C. Gomez-Laberge, G. Kreiman and R.T. Born. *Corticocortical feedback increases the spatial extent of normalization.* Frontiers in Systems Neurosciences, vol. 8, no. 105, 2014. (Cited on page 136.)

[Neumann 2001] H. Neumann and E. Mingolla. *Computational neural models of spatial integration in perceptual grouping.* In T.F.Shipley & P.J. Kellman, editor, From Fragments to Objects: Grouping and Segmentation in Vision, pages 353–400. Amsterdam: Elsevier, 2001. (Cited on pages 131 and 142.)

[Ni 2011] Z. Ni, C. Pacoret, R. Benosman, S. Ieng and S. Régnier. *Asynchronous event-based high speed vision for microparticle tracking.* Journal of Microscopy, vol. 245, no. 3, pages 236–244, November 2011. (Cited on page 125.)

[Nishimoto 2011] Shinji Nishimoto and Jack L. Gallant. *A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies.* The Journal of Neuroscience, vol. 31, no. 41, pages 14551–14564, 2011. (Cited on pages 43, 136, 137 and 142.)

[Noest 1993] A.J. Noest and A.V. Van Den Berg. *The role of early mechanisms in motion transparency and coherence.* Spatial Vision, vol. 7, no. 2, pages 125–147, 1993. (Cited on page 45.)

[Noh 2015] H. Noh, S. Hong and B. Han. *Learning deconvolution network for semantic segmentation.* In iccv, pages 1520–1528, Santiago, Chile, December 2015. (Cited on page 130.)

[Nothdurft 1991] H.C. Nothdurft. *Texture segmentation and pop-out from orientation contrast.* Vision Research, vol. 31, no. 6, pages 1073–1078, 1991. (Cited on page 127.)

[Nowlan 1994] S.J. Nowlan and T.J. Sejnowski. *Filter selection model for motion segmentation and velocity integration.* J. Opt. Soc. Am. A, vol. 11, no. 12, pages 3177–3199, 1994. (Cited on pages 44, 117, 138, 139, 141 and 142.)

[Odermatt 2012] Benjamin Odermatt, Anton Nikolaev and Leon Lagnado. *Encoding of Luminance and Contrast by Linear and Nonlinear Synapses in the Retina.* Neuron, vol. 73, no. 4, pages 758 – 773, 2012. (Cited on page 122.)

[O'Herron 2011] P. O'Herron and R. von der Heydt. *Representation of object continuity in the visual cortex.* J. of Vision, vol. 11, no. 2, pages 12, 1–9, 2011. (Cited on page 128.)

[Ohshiro 2011] Tomokazu Ohshiro, Dora E. Angelaki and Gregory C. DeAngelis. *A normalization model of multisensory integration.* Nature Neuroscience, vol. 14, no. 6, pages 775–782, May 2011. (Cited on pages 137 and 142.)

[Orban 2008] Guy A. Orban. *Higher Order Visual Processing in Macaque Extrastriate Cortex.* Physiological Reviews, vol. 88, no. 1, pages 59–89, 2008. (Cited on pages 63, 64, 105, 108, 115 and 135.)

[Orchard 2015] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor and Ryad Benosman. *HFirst: A Temporal Approach to Object Recognition*. pami, vol. 37, no. 10, October 2015. (Cited on page 125.)

[Pack 2001] C.C. Pack and R.T. Born. *Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain*. Nature, vol. 409, pages 1040–1042, February 2001. (Cited on pages 24, 51, 53 and 98.)

[Pack 2004] C.C. Pack, A.J. Gartland and R.T. Born. *Integration of Contour and Terminator Signals in Visual Area MT of Alert Macaque*. The Journal of Neuroscience, vol. 24, no. 13, pages 3268—3280, 2004. (Cited on pages 24 and 25.)

[Pack 2008] C.C. Pack and R.T. Born. *Cortical Mechanisms for the Integration of Visual Motion*. In Richard H. Masland, Thomas D. Albright, Thomas D. Albright, Richard H. Masland, Peter Dallos, Donata Oertel, Stuart Firestein, Gary K. Beauchamp, M. Catherine Bushnell, Allan I. Basbaum, Jon H. Kaas and Esther P. Gardner, editors, The Senses: A Comprehensive Reference, pages 189 – 218. Academic Press, New York, 2008. (Cited on pages 24, 36, 44, 45, 135 and 142.)

[Pal 1993] N.R. Pal and S.K. Pal. *A review of image segmentation techniques*. Pattern Recognition, vol. 26, no. 9, pages 1277–1294, 1993. (Cited on pages 126 and 130.)

[Palmer 1999] S. Palmer. Vision science : Photons to phenomenology. MIT Press, 1999. (Cited on page 4.)

[Pandarinath 2010] C. Pandarinath, J.D. Victor and S. Nirenberg. *Symmetry Breakdown in the ON and OFF Pathways of the Retina at Night: Functional Implications*. The Journal of neuroscience, vol. 30, no. 30, pages 10006–10014, 2010. (Cited on page 125.)

[Paninski 2004] Liam Paninski. *Maximum likelihood estimation of cascade point-process neural encoding models*. Network: Computation in Neural Systems, vol. 15, no. 4, pages 243–262, 2004. (Cited on page 48.)

[Parent 1989] P. Parent and S.W. Zucker. *Trace inference, curvature consistency, and curve detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 8, pages 823–839, 1989. (Cited on pages 131 and 141.)

[Paris 2009] S. Paris, P. Kornprobst, J. Tumblin and F. Durand. *Bilateral Filtering: Theory and Applications*. Foundations and Trends in Computer Graphics and Vision, vol. 4, no. 1, 2009. (Cited on page 49.)

[Perrinet 2004] L.U. Perrinet, M. Samuelides and S. Thorpe. *Coding static natural images using spiking event times: do neuron cooperate?* IEEE Transactions in Neural Networks and Learning Systems, vol. 15, no. 5, pages 1164–1175, 2004. (Cited on page 112.)

[Perrinet 2012] Laurent U. Perrinet and Guillaume S. Masson. *Motion-Based Prediction Is Sufficient to Solve the Aperture Problem.* Neural Computation, vol. 24, no. 10, pages 2726–2750, 2012. (Cited on pages 109 and 143.)

[Perrinet 2015] L.U. Perrinet. Biologically inspired computer vision, chapter Sparse models for computer vision, pages 319–346. Number 14. Wiley, 2015. (Cited on page 112.)

[Perrone 2008] J.A. Perrone and R.J. Krauzlis. *Spatial integration by MT pattern neurons: a closer look at pattern-to-component effects and the role of speed tuning.* Journal of Vision, vol. 8, no. 9, pages 1–14, 2008. (Cited on pages 39, 44, 83, 84 and 98.)

[Perrone 2012] John A. Perrone. *A neural-based code for computing image velocity from small sets of middle temporal (MT/V5) neuron inputs.* Journal of Vision, vol. 12, no. 8, 2012. (Cited on pages 117, 136, 141 and 142.)

[Pessoa 1998] L. Pessoa, E. Thompson and A. Noë. *Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception.* Behavioral and brain sciences, vol. 21, pages 723–802, 1998. (Cited on page 129.)

[Peterhans 1991] E. Peterhans and R. von der Heydt. *Subjective contours: bridging the gap between psychophysics and physiology.* Trends in Neurosciences, vol. 14, no. 3, pages 112–119, 1991. (Cited on pages 107 and 142.)

[Peterson 2008] Mary A. Peterson and Elizabeth Salvagio. *Inhibitory competition in figure-ground perception: Context and convexity.* Journal of Vision, vol. 8, no. 16, 2008. (Cited on pages 127 and 142.)

[Petitot 2003] J. Petitot. *An introduction to the Mumford-Shah segmentation model.* Journal of Physiology - Paris, vol. 97, pages 335–342, 2003. (Cited on page 131.)

[Petrou 2008] Maria Petrou and Anil Bharat. Next generation artificial vision systems: Reverse engineering the human visual system. Artech House Series Bioinformatics & Biomedical Imaging, 2008. (Cited on pages 103 and 113.)

[Piech 2013] V. Piech, W. Li, G.N. Reeke and C.D. Gilbert. *Network model of top-down influences on local gain and contextural interactions in visual cortex.* Proceedings of the National Academy of Sciences - USA, vol. 110, no. 43, pages 4108–4117, 2013. (Cited on page 107.)

[Pillow 2008] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky and Eero P Simoncelli. *Spatio-temporal correlations and visual signalling in a complete neuronal population.* Nature, vol. 454, no. 7207, pages 995–999, 2008. (Cited on pages 122 and 142.)

[Pinto 2009] N. Pinto, D. Doukhan, J.J. DiCarlo and D.D. Cox. *A high-throughput screening approach to discovering good forms of biologically inspired visual representation*. PLoS Comput Biol, vol. 5, no. 11, 2009. (Cited on pages 116 and 144.)

[Poggio 1985] Tomaso Poggio, Vincent Torre and Christof Koch. *Computational vision and regularization theory*. Nature, vol. 317, no. 6035, pages 314–319, September 1985. (Cited on page 127.)

[Poggio 2011] Tomaso Poggio and Gadi Geiger. *Werner Reichardt: the man and his scientific legacy*. MIT-CSAIL-TR-2011-011, CBCL-297, November 2011. (Cited on page 16.)

[Poggio 2012] T Poggio. *The Levels of Understanding framework, revised*. Perception, vol. 41, no. 9, pages 1017–1023, 2012. (Cited on page 112.)

[Pomplun 2012] Marc Pomplun and Junichi Suzuki, editors. Developing and applying biologically-inspired vision systems: Interdisciplinary concepts. IGI Global, 2012. (Cited on page 103.)

[Ponce-Alvarez 2013] Adrián Ponce-Alvarez, Alexander Thiele, Thomas D. Albright, Gene R. Stoner and Gustavo Deco. *Stimulus-dependent variability and noise correlations in cortical MT neurons*. Proceedings of the National Academy of Sciences, vol. 110, no. 32, pages 13162–13167, 2013. (Cited on page 84.)

[Poort 2012] J. Poort, F. Raudies, A. Wannig, V.A. Lamme, Neumann H. and P.R. Roelfsema. *The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex*. Neuron, vol. 108, no. 5, pages 1392–1402, 2012. (Cited on page 142.)

[Portelli 2014] Geoffrey Portelli, John Barrett, Evelyne Sernagor, Timothée Masquelier and Pierre Kornprobst. *The wave of first spikes provides robust spatial cues for retinal information processing*. Research Report RR-8559, INRIA, July 2014. (Cited on page 125.)

[Posch 2011] Christoph Posch, Daniel Matolin and Rainer Wohlgenannt. *A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS*. IEEE Journal of Solid-State Circuits, vol. 46, no. 1, pages 259–275, 2011. (Cited on pages 125 and 144.)

[Pouget 1998] Alexandre Pouget, Kechen Zhang, Sophie Deneve and Peter E. Latham. *Statistically Efficient Estimation Using Population Coding*. Neural Comput., vol. 10, no. 2, pages 373–401, February 1998. (Cited on pages 49 and 77.)

[Pouget 2000] A. Pouget, P. Dayan and R. Zemel. *Information processing with population codes*. Nature Reviews Neuroscience, vol. 1, no. 2, pages 125–132, 2000. (Cited on page 144.)

[Pouget 2003] Alexandre Pouget, Peter Dayan and Richard S. Zemel. *INFERENCE AND COMPUTATION WITH POPULATION CODES.* Annual Review of Neuroscience, vol. 26, no. 1, pages 381–410, 2003. PMID: 12704222. (Cited on page 112.)

[Pouget 2013] A. Pouget, J.M. Beck, W.J. Ma and P.E. Latham. *Probabilistic brains: knowns and unknowns.* Nature Neuroscience, vol. 16, no. 9, pages 1170–1178, 2013. (Cited on page 112.)

[Priebe 2003] Nicholas Priebe, Carlos Cassanello and Stephen Lisberger. *The neural representation of speed in macaque area MT/V5.* Journal of Neuroscience, vol. 23, no. 13, pages 5650–5661, July 2003. (Cited on pages 65 and 136.)

[Priebe 2006] N.J. Priebe, S.G. Lisberger and A.J. Movshon. *Tuning for Spatiotemporal Frequency and Speed in Directionally Selective Neurons of Macaque Striate Cortex.* The Journal of Neuroscience, vol. 26, no. 11, pages 2941—2950, 2006. (Cited on pages 107 and 142.)

[Pylyshyn 1999] Z. Pylyshyn. *Is vision continuous with cognition? The case for cognitive impenetrability of visual perception.* Behavioral and brain sciences, vol. 22, no. 3, pages 341–365, 1999. (Cited on page 118.)

[Qian 1994] N. Qian, R.A. Andersen and E.H. Adelson. *Transparent motion perception as detection of unbalanced motion signals. III. Modeling.* The journal of Neuroscience, vol. 14, no. 12, pages 7381–7392, December 1994. (Cited on page 84.)

[Qian 1997] N. Qian and R. A. Andersen. *A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena.* Vision Research, vol. 37, pages 1683–1698, 1997. (Cited on pages 137 and 142.)

[Rad 2011] Kamiar Rahnama Rad and Liam Paninski. *Information Rates and Optimal Decoding in Large Neural Populations.* In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira and Kilian Q. Weinberger, editors, NIPS, pages 846–854, 2011. (Cited on pages 49 and 76.)

[Raijmakers 1996] Maartje E. J. Raijmakers, Han L. H. van der Maas and Peter C. M. Molenaar. *Numerical bifurcation analysis of distance-dependent on-center off-surround shunting neural networks.* Biological Cybernetics, vol. 75, no. 6, pages 495–507, 1996. (Cited on page 85.)

[Rankin 2011] J. Rankin, E. Tlapale, R. Veltz, P. Kornprobst and O. Faugeras. *Multistability and bifurcations in a model of motion perception.* In Developments in Dynamical Systems Arising from the Biosciences, March 2011. (Cited on page 86.)

[Rankin 2013] James Rankin, Emilien Tlapale, Romain Veltz, Olivier Faugeras and Pierre Kornprobst. *Bifurcation analysis applied to a model of motion integration with a multistable stimulus.* Journal of Computational

Neuroscience, vol. 34, no. 1, pages 103–124, 2013. 10.1007/s10827-012-0409-5. (Cited on page 30.)

[Rankin 2014] James Rankin, Andrew I. Meso, Guillaume S. Masson, Olivier Faugeras and Pierre Kornprobst. *Bifurcation Study of a Neural Fields Competition Model with an Application to Perceptual Switching in Motion Integration.* Journal of Computational Neuroscience, vol. 36, no. 2, pages 193–213, 2014. (Cited on pages 86, 87, 98, 147 and 154.)

[Rao 1999] R.P. Rao and D.H. Ballard. *Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.* Nat Neurosci, vol. 2, no. 1, pages 79–87, 1999. (Cited on page 143.)

[Rasch 2013] Malte J. Rasch, Ming Chen, Si Wu, Haidong D. Lu and Anna W. Roe. *Quantitative inference of population response properties across eccentricity from motion-induced maps in macaque V1.* Journal of Neurophysiology, vol. 109, no. 5, pages 1233–1249, 2013. (Cited on pages 43 and 139.)

[Raudies 2010] Florian Raudies and Heiko Neumann. *A neural model of the temporal dynamics of figure–ground segregation in motion perception.* Neural Networks, vol. 23, no. 2, pages 160 – 176, 2010. (Cited on page 142.)

[Raudies 2011] Florian Raudies, Ennio Mingolla and Heiko Neumann. *A Model of Motion Transparency Processing with Local Center-Surround Interactions and Feedback.* Neural Computation, vol. 23, no. 11, pages 2868–2914, 2011. (Cited on pages 38, 40 and 138.)

[Raudies 2012] F. Raudies, E. Mingolla and H. Neumann. *Active Gaze Control Improves Optic Flow-Based Segmentation and Steering.* PLoS ONE, vol. 7, no. 6, pages 1–19, 06 2012. (Cited on page 133.)

[Reichardt 1961] W. Reichardt. *Autocorrelation, a principle for the evaluation of sensory information by the central nervous system.* In W. A. Rosenblith, editor, Principles of Sensory Communications, pages 303–317. John Wiley, New York, 1961. (Cited on pages 15 and 31.)

[Ren 2006] X. Ren, C.C. Fowlkes and J. Malik. *Figure/Ground Assignment in Natural Images.* In Aleš Leonardis, Horst Bischof and Axel Pinz, editors, Computer Vision – ECCV 2006, volume 3952 of *Lecture Notes in Computer Science*, pages 614–627. Springer Berlin Heidelberg, 2006. (Cited on pages 132, 141 and 142.)

[Reynaud 2012] A. Reynaud, G.S. Masson and F. Chavane. *Dynamics of local input normalization result from balanced short- and long-range intracortical interactions in area V1.* Journal of Neuroscience, vol. 32, pages 12558–12569, 2012. (Cited on pages 110, 111 and 136.)

[Reynolds 2000] J.H. Reynolds, T. Pasternak and R. Desimone. *Attention increases sensitivity of V4 neurons.* Neuron, vol. 26, pages 703–714, 2000. (Cited on page 110.)

[Reynolds 2009] J.H. Reynolds and D.J. Heeger. *The normalisation model of attention*. Neuron, vol. 61, pages 168–185, 2009. (Cited on pages 110 and 114.)

[Rodman 1987] H.R. Rodman and T.D. Albright. *Coding of visual stimulus velocity in area MT of the macaque*. Vision Research, vol. 27, no. 12, pages 2035–2048, 1987. (Cited on page 107.)

[Rodriguez Sanchez 2012] A. J. Rodriguez Sanchez and J. K. Tsotsos. *The Roles of Endstopped and Curvature Tuned Computations in a Hierarchical Representation of 2D Shape*. PLoS ONE, vol. 7, no. 8, page e42058, 2012. (Cited on pages 132, 141 and 142.)

[Roelfsema 2000] Pieter R. Roelfsema, Victor A. F. Lamme and Henk Spekreijse. *The implementation of visual routines*. Vision Research, vol. 40, no. 10–12, pages 1385–1411, 2000. (Cited on pages 104 and 109.)

[Roelfsema 2002] P. R. Roelfsema, V. A. F. Lamme, H. Spekreijse and H. Bosch. *Figure/ground segregation in a recurrent network architecture*. J. of Cognitive Neuroscience, vol. 14, no. 4, pages 525–537, 2002. (Cited on pages 109 and 129.)

[Roelfsema 2005] P.R. Roelfsema. *Elemental operations in vision*. Trends in Cognitive Sciences, vol. 9, no. 5, pages 226–233, 2005. (Cited on pages 109 and 130.)

[Roelfsema 2006] Pieter R. Roelfsema. *CORTICAL ALGORITHMS FOR PERCEPTUAL GROUPING*. Annual Review of Neuroscience, vol. 29, no. 1, pages 203–227, 2006. PMID: 16776584. (Cited on page 127.)

[Roelfsema 2007] Pieter R. Roelfsema, Michiel Tolboom and Paul S. Khayat. *Different Processing Phases for Features, Figures, and Selective Attention in the Primary Visual Cortex*. Neuron, vol. 56, no. 5, pages 785 – 792, 2007. (Cited on page 129.)

[Roelfsema 2011] PieterR. Roelfsema and Roos Houtkamp. *Incremental grouping of image elements in vision*. Attention, Perception, & Psychophysics, vol. 73, no. 8, pages 2542–2572, 2011. (Cited on pages 127 and 129.)

[Rogister 2012] P. Rogister, R. Benosman and S. H. Ieng. *Asynchronous event-based binocular stereo matching*. IEEE Transactions in Neural Networks and Learning Systems, pages 347–353, 2012. (Cited on page 125.)

[Rolls 2010] E.T. Rolls and G. Deco. The noisy brain: stochastic dynamics as a principle of brain function. Oxford university press, 2010. (Cited on page 112.)

[Rousselet 2004] Guillaume Rousselet, Simon Thorpe and Michele Fabre-Thorpe. *How parallel is visual processing in the ventral path?* TRENDS in Cognitive Sciences, vol. 8, no. 8, pages 363–370, August 2004. (Cited on page 109.)

[Rucci 2015] M. Rucci and J.D. Victor. *The unsteady eye: an information-processing stage, not a bug.* Trends in Neurosciences, vol. 38, no. 4, pages 195–206, 2015. (Cited on page 106.)

[Russell 2008] Bryan C. Russell, Antonio Torralba, KevinP. Murphy and WilliamT. Freeman. *LabelMe: A Database and Web-Based Tool for Image Annotation.* International Journal of Computer Vision, vol. 77, no. 1-3, pages 157–173, 2008. (Cited on page 145.)

[Rust 2005] N.C. Rust, O. Schwartz, J.A. Movshon and E.P. Simoncelli. *Spatiotemporal elements of macaque V1 receptive fields.* Neuron, vol. 46, pages 945–956, 2005. (Cited on page 142.)

[Rust 2006] N.C. Rust, V. Mante, E.P. Simoncelli and J.A. Movshon. *How MT cells analyze the motion of visual patterns.* Nature Neuroscience, vol. 9, pages 1421–1431, 2006. (Cited on pages 23, 43, 44, 45, 46, 48, 49, 75, 83, 84, 98, 117, 136, 137, 139 and 142.)

[Sato 2012] T.K. Sato, I. Nauhaus and M. Carandini. *Traveling waves in visual cortex.* Neuron, vol. 75, pages 218–229, 2012. (Cited on page 111.)

[Sceniak 1999] Michael P. Sceniak, Dario L. Ringach, Michael J. Hawken and Robert Shapley. *Contrast's effect on spatial summation by macaque V1 neurons.* Nature Neuroscience, vol. 2, no. 8, pages 733—739, August 1999. (Cited on page 65.)

[Scheirer 2014] W.J. Scheirer, S.E. Anthony, K. Nakayama and D.D. Cox. *Perceptual Annotation: Measuring Human Vision to Improve Computer Vision.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 36, no. 8, pages 1679–1686, 2014. (Cited on page 146.)

[Scherzer 2000] O. Scherzer and J. Weickert. *Relations between regularization and diffusion filtering.* Journal of Mathematical Imaging and Vision, vol. 12, no. 1, pages 43–63, February 2000. (Cited on page 139.)

[Scholte 2008] H.S. Scholte, J. Jolij, J.J. Fahrenfort and V.A.F. Lamme. *Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging.* J. of Cognitive Neuroscience, vol. 20, no. 11, pages 2097–2109, 2008. (Cited on pages 127 and 132.)

[Sejnowski 2014] Terrence J Sejnowski, Patricia S Churchland and J Anthony Movshon. *Putting big data to good use in neuroscience.* Nature Neuroscience, vol. 17, no. 11, pages 1440–1441, November 2014. (Cited on page 6.)

[Self 2013] M. W. Self, T. van Kerkoerle, H. Super and P. R. Roelfsema. *Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation.* Current Biology, vol. 23, no. 21, pages 2121 – 2129, 2013. (Cited on pages 137 and 142.)

[Sellent 2011] A. Sellent, M. Eisemann, B. Goldlucke, D. Cremers and M. Magnor. *Motion Field Estimation from Alternate Exposure Images*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 8, pages 1577–1589, 2011. (Cited on page 140.)

[Serre 2007a] T. Serre, A. Oliva and T. Poggio. *A Feedforward Architecture Accounts for Rapid Categorization*. Proceedings of the National Academy of Sciences (PNAS), vol. 104, no. 15, pages 6424–6429, 2007. (Cited on page 105.)

[Serre 2007b] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio. *Robust Object Recognition with Cortex-Like Mechanisms*. IEEE PAMI, vol. 29, no. 3, pages 411–426, March 2007. (Cited on page 105.)

[Shadlen 1999] M.N. Shadlen and J.A. Movshon. *Synchrony unbound: a critical evaluation of the temporal binding hypothesis*. Neuron, vol. 24, no. 1, pages 67–77, 1999. (Cited on page 112.)

[Shamir 2014] M. Shamir. *Emerging principles of population coding: in search for the neural code*. Current Opinion in Neurobiology, vol. 25, pages 140–148, 2014. (Cited on page 112.)

[Shapley 1984] R. Shapley and C. Enroth-Cugell. *Visual adaptation and retinal gain controls*. Progress in retinal research, vol. 3, pages 263–346, 1984. (Cited on pages 121 and 142.)

[Shapley 1990] R.B. Shapley. *Visual sensitivity and parallel retinocortical channels*. Annual Review of Psychology, vol. 41, pages 635–658, 1990. (Cited on page 121.)

[Shapley 2003] R.B. Shapley, M. Hawken and D.L. Ringach. *Dynamics of orientation selectivity in the primate visual cortex and the importance of cortical inhibition*. Neuron, vol. 38, no. 5, pages 689–699, 2003. (Cited on page 112.)

[Sharpee 2006] T.O. Sharpee, H. Sugihara, A.V. Kurgansky, S.P. Rebrik, M.P. Stryker and K.D. Miller. *Adaptive filtering enhances information transmission in visual cortex*. Nature, vol. 439, pages 936–942, 2006. (Cited on page 64.)

[Sheperd 2010] G.M. Sheperd and S. Grillner, editors. Handbook of brain microcircuits. Oxford University Press, 2010. (Cited on page 113.)

[Shi 2000] J. S. Shi and J. Malik. *Normalized Cuts and Image Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pages 888–905, 2000. (Cited on page 130.)

[Sigman 2001] A. Sigman, A. Cecchi, C.D. Gilbert and M.O. Magnaso. *On a common circle: Natural scenes and Gestalt rules*. PNAS, vol. 98, no. 4, pages 1935–1940, 2001. (Cited on pages 131 and 142.)

[Silies 2014] M. Silies, D.M. Gohl and T.R Clandinin. *Motion-Detecting Circuits in Flies: Coming into View.* Annual Review of Neuroscience, vol. 37, no. 1, pages 307–327, 2014. PMID: 25032498. (Cited on page 135.)

[Simoncelli 1991] Eero Simoncelli and Edward H. Adelson. *Computing Optical Flow Distributions Using Spatio-temporal Filters.* Technical report, MIT Media Lab Vision and Modeling, Tech. Rep, 1991. (Cited on pages 31 and 43.)

[Simoncelli 1993] Eero P. Simoncelli. *Course-to-fine Estimation of Visual Motion.* In IEEE Eighth Workshop on Image and Multidimensional Signal Processing, 1993. (Cited on page 50.)

[Simoncelli 1998] E.P. Simoncelli and D.J. Heeger. *A Model of Neuronal Responses in Visual Area MT.* Vision Research, vol. 38, pages 743–761, 1998. (Cited on pages 7, 30, 35, 44, 45, 46, 48, 49, 71, 74, 83, 136, 137, 139, 140 and 142.)

[Simoncelli 2001] E.P. Simoncelli and B.A. Olshausen. *Natural IMAGE STATISTICS AND NEURAL REPRESENTATION.* Annual Review of Neuroscience, vol. 24, no. 1, pages 1193–1216, 2001. (Cited on page 144.)

[Simoncini 2012] C. Simoncini, L.U. Perrinet, A. Montagnini, P. Mamassian and G.S. Masson. *More is not always: adaptive gain control explains dissociation between perception and action.* Nature Neuroscience, vol. 15, no. 11, pages 1586–1603, 2012. (Cited on pages 136 and 137.)

[Sincich 2005] Lawrence C. Sincich and Jonathan C. Horton. *THE CIRCUITRY OF V1 AND V2: Integration of Color, Form, and Motion.* Annual Review of Neuroscience, vol. 28, no. 1, pages 303–326, 2005. PMID: 16022598. (Cited on page 43.)

[Singer 1999] W. Singer. *Neuronal synchrony: a versatile code for the definition of relations?* Neuron, vol. 24, no. 1, pages 49–65, 1999. (Cited on page 111.)

[Skottun 1999] Bernt Christian Skottun. *Neuronal responses to plaids.* Vision Research, vol. 39, no. 12, pages 2151 – 2156, 1999. (Cited on page 45.)

[Smith 2005] M. Smith, N. Majaj and A. Movshon. *Dynamics of motion signaling by neurons in macaque area MT.* Nature Neuroscience, vol. 8, no. 2, pages 220–228, February 2005. (Cited on pages 25, 96 and 98.)

[Smolyanskaya 2013] Alexandra Smolyanskaya, Douglas Andrew Ruff and Richard T. Born. *Joint tuning for direction of motion and binocular disparity in macaque MT is largely separable.* Journal of Neurophysiology, 2013. (Cited on pages 137 and 142.)

[Snowden 1991] R. J. Snowden, S. Treue, R. G. Erickson and R. A. Andersen. *The response of area MT and V1 neurons to transparent motion.* The Journal of Neuroscience, vol. 11, no. 9, pages 2768–2785, 1991. (Cited on pages 22, 26, 27 and 137.)

[Solari 2015] F. Solari, M. Chessa, K. Medathati and P. Kornprobst. *What can we expect from a V1-MT feedforward architecture for optical flow estimation?* Signal Processing: Image Communication, vol. 39, no. B, pages 342–354, 2015. (Cited on pages 117, 139, 140, 141, 142 and 147.)

[Squire 2013] R.F. Squire, B. Noudoost, R.J. Schafer and T. Moore. *Prefrontal contributions to visual selective attention.* Annual Review of Neuroscience, vol. 36, pages 451–466, 2013. (Cited on page 110.)

[Stein 2009] Andrew N. Stein and Martial Hebert. *Occlusion Boundaries from Motion: Low-Level Detection and Mid-Level Reasoning.* International Journal of Computer Vision, vol. 82, no. 3, pages 325–357, 2009. (Cited on page 131.)

[Stoner 1990] G.R. Stoner, T.D Albright and V.S. Ramachandran. *Transparency and coherence in human motion perception.* Nature, vol. 344, no. 6262, March 1990. (Cited on pages 18 and 45.)

[Stoner 1992] Gene R. Stoner and Thomas D. Albright. *Neural correlates of perceptual motion coherence.* Nature, vol. 358, pages 412–414, 1992. (Cited on page 137.)

[Sugita 1999] Y. Sugita. *Grouping of image fragments in primary visual cortex.* Nature, vol. 401, no. 6750, pages 269–272, 1999. (Cited on page 107.)

[Sun 2010] D. Sun, S. Roth, T.U. Darmstadt and M.J. Black. *Secrets of Optical Flow Estimation and Their Principles.* In Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2010. (Cited on page 138.)

[Sundberg 2011] Patrik Sundberg, Thomas Brox, Michael Maire, Pablo Arbeláez and Jitendra Malik. *Occlusion Boundary Detection and Figure/Ground Assignment from Optical Flow.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. (Cited on pages 131 and 132.)

[Thiele 2001a] A. Thiele, K.R. Dobkins and T.D. Albright. *Neural correlates of chromatic motion perception.* Neuron, vol. 32, no. 351-358, 2001. (Cited on pages 108 and 115.)

[Thiele 2001b] A. Thiele and J.A. Perrone. *Speed skills: measuring the visual speed analyzing properties of primate MT neurons.* Nature Neuroscience, vol. 4, no. 5, pages 526–532, 2001. (Cited on page 142.)

[Thoreson 2012] W.B. Thoreson and S.C. Mangel. *Lateral interactions in the outer retina.* Prog Retin Eye Res., vol. 31, no. 5, pages 407–441, 2012. (Cited on pages 121 and 142.)

[Thorpe 2001] S.J. Thorpe, A. Delorme and R. Van Rullen. *Spike-based strategies for rapid processing.* Neural Networks, vol. 14, no. 6-7, pages 715–725, 2001. (Cited on page 112.)

[Thorpe 2009] S.J. Thorpe. *The speed of categorization in the human visual system.* Neuron, vol. 62, no. 2, pages 168–170, 2009. (Cited on page 109.)

[Tlapale 2010] Emilien Tlapale, Guillaume S. Masson and Pierre Kornprobst. *Modelling the dynamics of motion integration with a new luminance-gated diffusion mechanism.* Vision Research, vol. 50, no. 17, pages 1676–1692, August 2010. (Cited on pages 44, 109, 115, 117, 136, 138, 141, 142, 143 and 147.)

[Tlapale 2011a] Emilien Tlapale. *Modelling the dynamics of contextual motion integration in the primate.* PhD thesis, Université Nice Sophia Antipolis, January 2011. (Cited on page 20.)

[Tlapale 2011b] Emilien Tlapale, Pierre Kornprobst, Guillaume S. Masson and Olivier Faugeras. *A Neural Field Model for Motion Estimation.* In Springer Verlag, editor, Mathematical Image Processing, volume 5 of *Springer Proceedings in Mathematics*, pages 159–180, 2011. (Cited on pages 30, 38, 40, 116, 139 and 147.)

[Treue 2000] S. Treue, K. Hol and H.J. Rauber. *Seeing multiple directions of motion – physiology and psychophysics.* Nature Neuroscience, vol. 3, no. 3, pages 270–276, 2000. (Cited on pages 27, 28, 39 and 84.)

[Tschechne 2014a] Stephan Tschechne and Heiko Neumann. *Hierarchical representation of shapes in visual cortex - from localized features to figural shape segregation.* Frontiers in Computational Neuroscience, vol. 8, no. 93, 2014. (Cited on pages 132, 141 and 142.)

[Tschechne 2014b] Stephan Tschechne, Roman Sailer and Heiko Neumann. *Bio-Inspired Optic Flow from Event-Based Neuromorphic Sensor Input.* Artificial Neural Networks in Pattern Recognition, vol. 8774, pages 171–182, 2014. (Cited on pages 125 and 141.)

[Tsotsos 1993] J.K. Tsotsos. Spatial vision in humans and robots, chapter An inhibitory beam for attentional selection, pages 313–331. Cambridge University Press, 1993. (Cited on page 109.)

[Tsotsos 1995] J.K. Tsotsos, S.M. Culhane, W.Y. Kei Wai, Y. Lai, N. Davis and F. Nuflo. *Modeling visual attention via selective tuning.* Artificial Intelligence, vol. 78, pages 507–545, 1995. (Cited on pages 109 and 110.)

[Tsotsos 2011] J. Tsotsos. A computational perspective on visual attention. The MIT Press, May 2011. (Cited on pages 110, 118 and 119.)

[Tsotsos 2014] John K. Tsotsos. *It's all about the constraints.* Current Biology, vol. 24, no. 18, pages 854–858, September 2014. (Cited on pages 103, 118 and 145.)

[Tsotsos 2015] John K. Tsotsos, Miguel P. Eckstein and Michael S. Landy. *Computational models of visual attention.* Vision Research, vol. 116, Part B, pages 93 – 94, 2015. Computational Models of Visual Attention. (Cited on pages 110 and 119.)

[Tsui 2010] James M. G. Tsui, J. Nicholas Hunter, Richard T. Born and Christopher C. Pack. *The Role of V1 Surround Suppression in MT Motion Integration*. Journal of Neurophysiology, vol. 103, no. 6, pages 3123–3138, 2010. (Cited on pages 136 and 139.)

[Turpin 2014] Andrew Turpin, David J. Lawson and Allison M. McKendrick. *PsyPad: A platform for visual psychophysics on the iPad*. Journal of Vision, vol. 14, no. 3, 2014. (Cited on page 145.)

[Ullman 1984] S. Ullman. *Visual routines*. Cognition, vol. 18, pages 97–159, 1984. (Cited on page 129.)

[Ullman 1995] S. Ullman. *Sequence seeking and counter streams A computational model for bidirectional information flow in the visual cortex.* Cerebral Cortex, vol. 5, pages 1–11, 1995. (Cited on page 128.)

[Ullman 2002] S. Ullman, M. Vidal-Naquet and E. Sali. *Visual features of intermediate complexity and their use in classification*. Nature Neuroscience, vol. 5, no. 7, pages 682–687, 2002. (Cited on page 131.)

[Ullman 2007] S. Ullman. *Object recognition and segmentation by a fragment-based hierarchy*. Trends in Cognitive Science, vol. 11, no. 2, pages 58–64, 2007. (Cited on page 131.)

[Ullman 2016] S. Ullman, L. Assif, E. Fataya and D. Harari. *Atoms of recognition in human and computer vision*. Proceedings of the National Academy of Sciences - USA, vol. 113, no. 10, pages 2744–2749, 2016. (Cited on page 118.)

[Ungerleider 1982] L.G. Ungerleider and M. Mishkin. Two cortical visual systems, pages 549–586. MIT Press, 1982. (Cited on page 104.)

[Ungerleider 1994] Leslie G. Ungerleider and James V. Haxby. *'What' and 'where' in the human brain*. Current Opinion in Neurobiology, vol. 4, no. 2, pages 157–165, 1994. (Cited on page 104.)

[Valeiras 2016] David Reverter Valeiras, Garrick Orchard, Sio-Hoi Ieng and Ryad Benosman. *Neuromorphic Event-Based 3D Pose Estimation*. Frontiers in Neuroscience, vol. 1, no. 9, January 2016. (Cited on page 125.)

[Van Essen 2003] D. C. Van Essen. *Organization of visual areas in macaque and human cerebral cortex*. In L Chapula and J Werner, editors, The Visual Neurosciences. MIT Press, 2003. (Cited on page 104.)

[Van Santen 1984] JPH Van Santen and G. Sperling. *Temporal covariance model of human motion perception*. Journal of the Optical Society of America A, vol. 1, no. 5, pages 451–473, 1984. (Cited on page 16.)

[Van Santen 1985] J.P.H. Van Santen and G. Sperling. *Elaborated Reichardt detectors*. Journal of the Optical Society of America A, vol. 2, no. 2, pages 300–320, 1985. (Cited on pages 31 and 32.)

[VanRullen 2002] R. VanRullen and S. J. Thorpe. *Surfing a spike wave down the ventral stream*. Vision Research, vol. 42, pages 2593–2615, 2002. (Cited on pages 125 and 141.)

[Verri 1987] A. Verri and T. Poggio. *Against quantitative optical flow*. In Proceedings First International Conference on Computer Vision, pages 171–180. IEEE Computer Society, 1987. (Cited on page 133.)

[Verri 1989] A. Verri and T. Poggio. *Motion field and optical flow: qualitative properties*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 5, pages 490–498, 1989. (Cited on page 133.)

[Vetter 2014] P. Vetter and A. Newen. *Varieties of cognitive penetration in visual perception*. Conscious Cognition, vol. 27, pages 62–75, 2014. (Cited on page 114.)

[Vinje 2000] W.E. Vinje and J. L. Gallant. *Sparse coding and decorrelation in primary visual cortex during natural vision*. Science, vol. 287, pages 1273–1276, 2000. (Cited on page 136.)

[von der Heydt 2015] R. von der Heydt. *Figure-ground organisation and the emergence of proto-objects in the visual cortex*. Frontiers in Psychology, vol. 6, no. 1695, 2015. (Cited on page 107.)

[von der Malsburg 1981] Christoph von der Malsburg. *The Correlation Theory of Brain Function*. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, 1981. (Cited on page 111.)

[Von der Malsburg 1999] C. Von der Malsburg. *The what and why of binding: the modeler's perspective*. Neuron, vol. 24, no. 1, pages 95–104, 1999. (Cited on page 111.)

[Vondrick 2013] Carl Vondrick, Donald Patterson and Deva Ramanan. *Efficiently Scaling up Crowdsourced Video Annotation*. International Journal of Computer Vision, vol. 101, no. 1, pages 184–204, 2013. (Cited on page 145.)

[Wallace 2005] J.M. Wallace, L.S. Stone and G.S. Masson. *Object Motion Computation for the Initiation of Smooth Pursuit Eye Movements in Humans*. Journal of Neurophysiology, vol. 93, no. 4, pages 2279–2293, 2005. (Cited on pages 19 and 20.)

[Wallach 1935] H. Wallach. *Über visuell wahrgenommene Bewegungsrichtung*. Psychological Research, vol. 20, no. 1, pages 325–380, 1935. (Cited on pages 16, 17 and 20.)

[Wang 1997] Ruye Wang. *A Network Model of Motion Processing in Area MT of Primates*. Journal of Computational Neuroscience, vol. 4, no. 4, pages 287–308, 1997. (Cited on page 84.)

[Warren 2012] W.H. Warren. *Does this computational theory solve the right problem? Marr, Gibson and the goal of vision*. Perception, vol. 41, no. 9, pages 1053–1060, 2012. (Cited on page 118.)

[Webb 2003] B.S. Webb, C.J. Tinsley, N.E. Barraclough, A. Parker and A.M. Derrington. *Gain control from beyond the classical receptive field in primate visual cortex.* Visual Neuroscience, vol. 20, no. 3, pages 221–230, 2003. (Cited on page 111.)

[Wedel 2009] A. Wedel, D. Cremers, T. Pock and H. Bischof. *Structure- and motion-adaptive regularization for high accuracy optic flow.* In Computer Vision, 2009 IEEE 12th International Conference on, pages 1663–1668, 2009. (Cited on page 140.)

[Weidenbacher 2009] U. Weidenbacher and H. Neumann. *Extraction of surface-related features in a recurrent model of V1-V2 interaction.* PLoS ONE, vol. 4, no. 6, 2009. (Cited on page 132.)

[Weiss 1998] Y. Weiss. *Bayesian motion estimation and segmentation.* PhD thesis, Massachusetts Institute of Technology, 1998. (Cited on page 7.)

[Willems 2011] Roel M Willems. *Re-Appreciating the Why of Cognition: 35 Years after Marr and Poggio.* Frontiers in Psychology, vol. 2, 2011. (Cited on page 115.)

[Williford 2013] J.R. Williford and R. von der Heydt. *Border-ownership coding.* Scholarpedia, vol. 8, no. 10, page 30040, 2013. (Cited on pages 128 and 129.)

[Wilson 1973] H.R. Wilson and J.D. Cowan. *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue.* Biological Cybernetics, vol. 13, no. 2, pages 55–80, September 1973. (Cited on page 85.)

[Wilson 1992] H.R. Wilson, V.P. Ferrera and C. Yo. *A psychophysically motivated model for two-dimensional motion perception.* Visual Neuroscience, vol. 9, no. 1, pages 79–97, July 1992. (Cited on page 17.)

[Witkin 1983] A.P. Witkin and J.M. Tenenbaum. Human and machine vision, chapter On the role of structure in vision, pages 481–543. Academic Press, 1983. (Cited on page 127.)

[Wohrer 2009] Adrien Wohrer and Pierre Kornprobst. *Virtual Retina : A biological retina model and simulator, with contrast gain control.* Journal of Computational Neuroscience, vol. 26, no. 2, page 219, 2009. DOI 10.1007/s10827-008-0108-4. (Cited on pages 123, 124, 141, 142, 143 and 147.)

[Wolfe 1991] W.J. Wolfe, D. Mathis, C. Anderson, J. Rothman, M. Gottler, G. Brady, R. Walker, G. Duane and G. Alaghband. *K-winner networks.* Neural Networks, IEEE Transactions on, vol. 2, no. 2, pages 310–315, 1991. (Cited on page 85.)

[Wolfe 2002] Jeremy M Wolfe, Aude Oliva, Todd S Horowitz, Serena J Butcher and Aline Bompas. *Segmentation of objects from backgrounds in visual*

search tasks. Vision Research, vol. 42, no. 28, pages 2985 – 3004, 2002. (Cited on pages 127 and 142.)

[Wuerger 1996] S. Wuerger, R. Shapley and N. Rubin. *"On the visually perceived direction of motion" by Hans Wallach: 60 years later*. Perception, vol. 25, pages 1317–1367, 1996. (Cited on page 17.)

[Xiao 1995] D. Xiao, S. Raiguel, V. Marcar, J. Koenderink and G. A. Orban. *Spatial Heterogeneity of Inhibitory Surrounds in the Middle Temporal Visual Area*. Proceedings of the National Academy of Sciences, vol. 92, no. 24, pages 11303–11306, 1995. (Cited on page 111.)

[Xiao 2015] Jianbo Xiao and Xin Huang. *Distributed and Dynamic Neural Encoding of Multiple Motion Directions of Transparently Moving Stimuli in Cortical Area MT.* The Journal of Neuroscience, vol. 35, no. 49, pages 16180–16198, December 2015. (Cited on pages 22, 23, 25, 26, 27, 29, 39, 83, 84, 88, 90, 92, 93, 94, 95, 96, 97, 98 and 99.)

[Xie 2002] Xiaohui Xie, Richard H R Hahnloser and H Sebastian Seung. *Selectively Grouping Neurons in Recurrent Networks of Lateral Inhibition*. Neural Computation, vol. 14, no. 11, pages 2627–2646, January 2002. (Cited on page 85.)

[Yabuta 292] N.H. Yabuta, A. Sawatari and E.M. Callaway. *Two functional channels from primary visual cortex to dorsal visual cortex*. 2001, vol. 297-301, 292. (Cited on page 108.)

[Yamins 2014] D.L.K. Yamins, H. Hong, C.F. Cadieu, E.A. Solomon, D. Seibert and J.J. DiCarlo. *Performance-optimized hierarchical models predict neural responses in higher visual cortex*. PNAS, vol. 111, no. 23, pages 8619–8624, 2014. (Cited on page 145.)

[Yamins 2016] D.L.K. Yamins and J.J. DiCarlo. *Using goal-driven deep learning models to understand sensory cortex*. Nature Neuroscience, vol. 19, no. 3, pages 356–365, 2016. (Cited on page 116.)

[Yang 2014] Z. Yang, D.J. Heeger, Blake R. and E. Seidemann. *Long-range traveling waves of activity triggered by local dichoptic stimulation in V1 of behaving monkeys*. Journal of Neurophysiology, 2014. (Cited on page 142.)

[Yarbus 1967] A. L. Yarbus. Eye movements and vision, chapter 1-7. Plenum Press, 1967. (Cited on page 132.)

[Yo 1992] C. Yo and H.R. Wilson. *Perceived direction of moving two-dimensional patterns depends on duration, contrast and eccentricity.* Vision Research, vol. 32, no. 1, pages 135–47, 1992. (Cited on page 16.)

[Yuille 1989] A.L. Yuille and N.M. Grzywacz. *A Winner-Take-All Mechanism Based on Presynaptic Inhibition Feedback*. Neural Computation, vol. 1, no. 3, pages 334–347, 1989. (Cited on page 85.)

[Zeki 1993] S. Zeki. A vision of the brain. Blackwell Scientific Publications, 1993. (Cited on page 107.)

[Zhang 2012] Yifeng Zhang, In-Jung Kim, Joshua R Sanes and Markus Meister. *The most numerous ganglion cell type of the mouse retina is a selective feature detector*. Proceedings of the National Academy of Sciences, vol. 109, no. 36, pages E2391–E2398, 2012. (Cited on page 114.)

[Zhou 2000] Hong Zhou, Howard S. Friedman and Rüdiger von der Heydt. *Coding of Border Ownership in Monkey Visual Cortex*. The Journal of Neuroscience, vol. 20, no. 17, pages 6594–6611, 2000. (Cited on pages 107, 128, 129 and 142.)

[Zhuo 2003] Y. Zhuo, T.G. Zhou, H.Y. Rao, J.J. Wang, M. Meng, M. Chen, C. Zhou and L. Chen. *Contributions of the visual ventral pathway to long-range apparent motion*. Science, vol. 299, no. 5605, page 417, 2003. (Cited on page 108.)

[Zucker 1981] S.W. Zucker. *Computer vision and human perception*. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, pages 1102–1116, 1981. (Cited on page 118.)