PhD THESIS

prepared at
INRIA Sophia Antipolis

and presented at the
University of Nice-Sophia Antipolis
Graduate School of Information and Communication Sciences

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF SCIENCE
Specialized in Control, Signal and Image Processing

Bio-Inspired Models for Motion Estimation and Analysis: Human action recognition and motion integration

María–José ESCOBAR

Defense date: November 27th, 2009

Adviser	Pierre KORNPROBST	INRIA Sophia Antipolis, France
Co-adviser	Thierry VIEVILLE	INRIA Sophia Antipolis, France
Reviewers	Martin GIESE	University Clinic Tübingen, Germany
	Heiko NEUMANN	Ulm University, Germany
Examiners	Olivier FAUGERAS	INRIA Sophia Antipolis, France
	Guillaume MASSON	Laboratoire DyVa, CNRS Marseille, France
	Pascal MAMASSIAN	Université Paris Descartes, France
	Javier RUIZ DEL SOLAR	Universidad de Chile, Chile

UNIVERSITÉ NICE-SOPHIA ANTIPOLIS - UFR Sciences

École Doctorale STIC
(Sciences et Technologies de l'Information et de la Communication)

THÈSE

pour obtenir le titre de
DOCTEUR EN SCIENCES
de l'UNIVERSITÉ de Nice-Sophia Antipolis
Discipline: Automatique, Traitement du Signal et des Images

Modèles bio-inspirés pour l'estimation et l'analyse de mouvement: Reconnaissance d'actions et intégration du mouvement

María–José ESCOBAR

Date prévue de soutenance, 27 novembre 2009

Composition du jury:

Directeur de thèse	Pierre KORNPROBST	INRIA Sophia Antipolis, France
Co-directeur de thèse	Thierry VIEVILLE	INRIA Sophia Antipolis, France
Rapporteurs	Martin GIESE	University Clinic Tübingen, Germany
	Heiko NEUMANN	Ulm University, Germany
Examinateurs	Olivier FAUGERAS	INRIA Sophia Antipolis, France
	Guillaume MASSON	Laboratoire DyVa, CNRS Marseille, France
	Pascal MAMASSIAN	Université Paris Descartes, France
	Javier RUIZ DEL SOLAR	Universidad de Chile, Chile

This work was partially supported by:
FACETS (EC IP project FP6-015879) and CONICYT Chile.

Contents

List of Figures

List of Tables

Abstract

This thesis addresses the study of the motion perception in mammals and how bio-inspired systems can be applied to real applications. The first part of this thesis relates how the visual information is processed in the mammal's brains and how motion estimation is usually modeled. Based on this analysis of the state of the art, we propose a feedforward V1-MT core architecture. This feedforward V1-MT core architecture will be a basis to study two different kinds of applications. The first application is human action recognition, which is still a challenging problem in the computer vision community. We show how our bio-inspired method can be successfully applied to this real application. Interestingly, we show how several computational properties inspired from motion processing in mammals, allow us to reach high quality results, which will be compared to latest reference results. The second application of the bio-inspired architecture proposed in this thesis, is to consider the problem of motion integration for the solution of the aperture problem. We investigate the role of delayed V1 surround suppression, and how the 2D information extracted through this mechanism can be integrated to propose a solution for the aperture problem. Finally, we highlight a variety of important issues in the determination of motion estimation and additionally we present many potential avenues for future research efforts.

Résumé

Cette thèse porte sur l'étude et la modélisation de la perception du mouvement chez le mammifère. Nous montrons comment un système bio-inspiré peut être appliqué dans le cadre d'une application réelle de vision par ordinateur, mais aussi comment il permet de mieux comprendre des phénomènes observés en neurosciences. La première partie de cette thèse étudie comment l'information visuelle est traitée chez le mammifère et comment l'estimation du mouvement est classiquement modelisée. A partir de cette analyse de l'état de l'art, nous avons proposé une architecture sequentielle générale, modélisant les aires corticales V1 et MT. Nous avons utilisé cette architecture pour étudier deux applications. La première application est la reconaissance d'actions dans les séquences d'images, problèmatique encore ouvert en vision par ordinateur. Nous montrons comment notre architecture bio-inspirée peut être appliquée avec succés dans le cadre de cette application réelle, en y apportant de nouvelles idées. En particulier, nous montrons comment la prise en compte de plusieurs propriétés du système visuel chez le mammifère nous permettent d'obtenir des résultats de haute qualité, comparables à ceux des approches les plus récentes. La deuxième application de l'architecture bio-inspirée proposée dans le cadre de cette thèse, est de chercher à comprendre la dynamique de l'intégration du mouvement. Pour cela, nous avons cherché à comprendre le rôle fonctionnel de la suppression du pourtour des neurones de V1. Notre modèle montre comment l'information 2D extraite à partir de ce mécanisme de suppression peut être intégrée dans la solution du problème d'ouverture. Enfin de nombreuse perspectives concluent ce travail, qui montrent combien l'étude de l'estimation de mouvement conserve encore de nombreuses problematiques.

CHAPTER 1

INTRODUCTION

"What does it mean to see? The plain man's answer would be, to know what is where by looking. In other words, vision is the process of discovering from images what is present in the world, and where it is."
– David Marr

1.1 UNDERSTANDING VISION

52%

52% is the percentage of the monkey's cortical area dedicated to vision. Just compare it with the other sensory functionalities! somatosensory: 10%, motor: 8%, auditory: 3%, olfactory: 1%. So what makes visual information so demanding?

David Marr, as a pioneer in computational neurosciences, stated that *vision* is not only the action of seeing, but also the action of processing what is present in the world and where it is. According to him, the study of vision must therefore not only include the study of isolated properties, but also the nature of the internal representations by which we capture this information and thus make it available as a basis for decision about our thoughts and actions. In other words, a duality between representation and information processing.

Understanding vision, requires the understanding of the visual system, which is one of the most ambitious project in science of the last 50 years. Get into the brain, connect electrodes, differentiate functional areas or measure the activity of population of neurons; all these efforts made following the same direction, which is, to better understand cortical processing from the functional level down to the neuronal level.

Thanks to these neurophysiological studies, connections between different brain areas have been established revealing the big complexity of the visual system, which is illustrated in Figure 1.1. Within visual brain areas, it has been possible to also identify those ones related to motion analysis, being the main ones: V1, MT and MST.

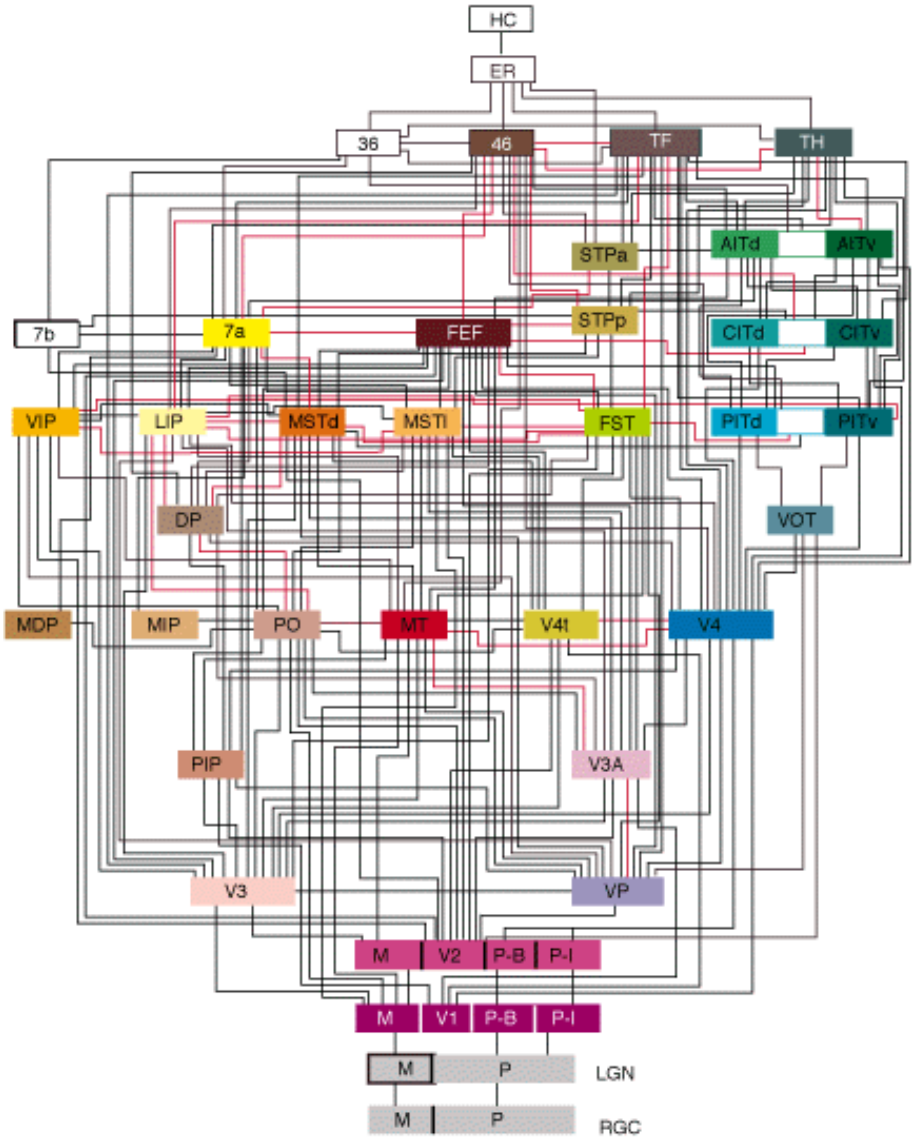


Figure 1.1: Circuit diagram of macaque visual areas (from Felleman and Van Essen (1991)).

Motion

Within the mechanisms related to vision, motion analysis is one of the most important. The visual detection of motion is crucial for survival of all but the very simplest creatures. Moving objects are likely to be a dangerous predator, or potential food, or a mate. Indeed, to produce signals in the absence of movement is a property of eyes quite high up the evolutionary scale. But motion is also crucial for a number of other tasks. For example, let us consider the case of that patient who suffered a stroke damaging her extrastiate region, thought to be involved in motion analysis. That patient was unable to appreciate the motion of objects, but also she had difficulties to pour tea into a cup because the fluid "seemed" frozen, or to follow a dialog because she could not follow the movements of the speaker's mouth. This illustrates the role

of motion in our daily activities, and even in our social interactions.

Motion is also crucial as soon as we consider vision application dealing with videos. Indeed, any real system needs to have a motion estimation at some stage. Some examples include robotics, autonomous navigation, weather forecast, video restoration, content retrieval, video surveillance, crowd analysis and action recognition. So there are strong needs to have some robust and precise algorithms to estimate motion in real scenes, and this justifies the very large literature in this domain done by the computer vision community. Interestingly, this problem remains challenging, and we can wonder whether new ideas coming from biology could improve the current performances.

Goal and methodology

In order to understand some functionalities of the visual processing, the objective of this thesis is to propose a bio-inspired feedforward model for motion analysis. Under our definition, the bio-inspiration term is assigned to models that either follow the brain hierarchical architecture in some sense, or model functionalities or operations found in real cell recordings, or implement analogue or spiking neurons. Bio-inspired models therefore can be useful to understand many properties of the visual system but just a few of them are proposed to deal with real sequences in order to experience *vision*.

In this thesis we are interested in motion processing. We studied the mechanisms involved in this task in the mammal brain and we proposed feedforward V1-MT models to lead with two classical topics related to motion analysis: human action recognition and motion integration.

Human action recognition

Human action recognition is the task of recognize the action performed in image sequences, assigning to each sequence an action label indicating the action taking place, e.g., walking, running, etc.

Initially, the analysis of human motion was studied by the psychologist Gunner Johansson in 1973 (Johansson (1973)). He produced extremely striking demonstrations of how little information is needed for seeing moving humans, and animals. Placing lights at the joints of someone's arms and legs, he or she becomes invisible in a dark room. As soon as the person moves, the lights are seen as a human figure, and even the gender can be identified. This type of stimulus, called *biological motion* or *point-light stimulus* is a highly complex motion pattern and interesting in many senses: it links the perception of motion with form (Pucel and Perret (2003); Michels et al. (2005); Hirai and Hiraki (2006)), it removes distractors such as background, clothing, etc.

But the real world is not seen as point-light stimuli, our visual system receives as

input real video sequences that must be fully interpreted to recognize patterns. And according to these patterns interact with the environment. Most of our social life and interaction with the environment comes from the recognition of members of our own specie, specially the actions that they are performing. As humans, we can easily recognize if a person is approaching to us walking, running, if someone is waving one hand, two hands, etc.

Automatizing this task has interesting applications in different domains including visual surveillance, video retrieval and human-computer interaction. This problem has been classically treated in the computer vision community, where the methods proposed do not pretend to be bio-inspired.

Currently, within all the methods proposed both in the computer vision community and in the computational neuroscience community, any of them is able to perform human action recognition in a variety of conditions and scenarios, motivating of this way the development of new approaches following new tracks. We asked whether the development of bio-inspired methodologies could bring new insights that have not been studied before.

In this thesis we developed two different architectures to convey human action recognition: We explore how actions can be represented from analog MT output or spike trains, how different center-surround interactions in MT neurons affect the performance of the human action recognition task, etc.

Motion integration

Motion integration mechanisms convey to the solution of the aperture problem. The aperture problem is a classical crossroad in visual neuroscience and many models have been proposed as solution. The aperture problem has called the attention of scientist along history, observing that the real movement of objects in the world is relative, and the only way to measure it is by reference to other objects.

Several mechanisms have been proposed to understand how the mammal visual system computes the real motion direction of objects. Within the mechanisms proposed we can cite: end-detector information coming from V2, spatial diffusion of non-ambiguous motion signals, feedbacks from upper layers, surround suppression, winner-take-all mechanisms, etc.

Inspired by the recent findings reported by Pack et al. (2004), in this thesis we asked whether the V1 surround-suppression can lead the solution of the aperture problem. The V1 surround suppression was implemented in the feedforward V1-MT architecture proposed in this thesis, and we explore its effect in the preferred direction of MT neurons, which have been reported as a mechanism for the solution of the aperture problem (Pack et al. (2004); Huang et al. (2007); Tlapale et al. (2008)).

1.2 Organization and main contribution

This thesis is organized in four parts (see Figure 1.2):

- Part I: *Motion Perception in Mammals* describes the studies related to how the visual motion information is processed in the mammal brain and how visual motion mechanisms have been classically modeled: on its detection and processing. We also show the feedforward V1-MT core architecture proposed and developed in this thesis.
- Part II: *Human Action Recognition* shows the state of the art of this problem together with two different proposals, created from the core model defined in Part I, to perform human action recognition.
- Part III shows how the V1 surround inhibition can be involved in the motion integration mechanism to solve the aperture problem.
- Finally, Part IV groups the conclusion, perpectives and publications associated to the work developed in this thesis.

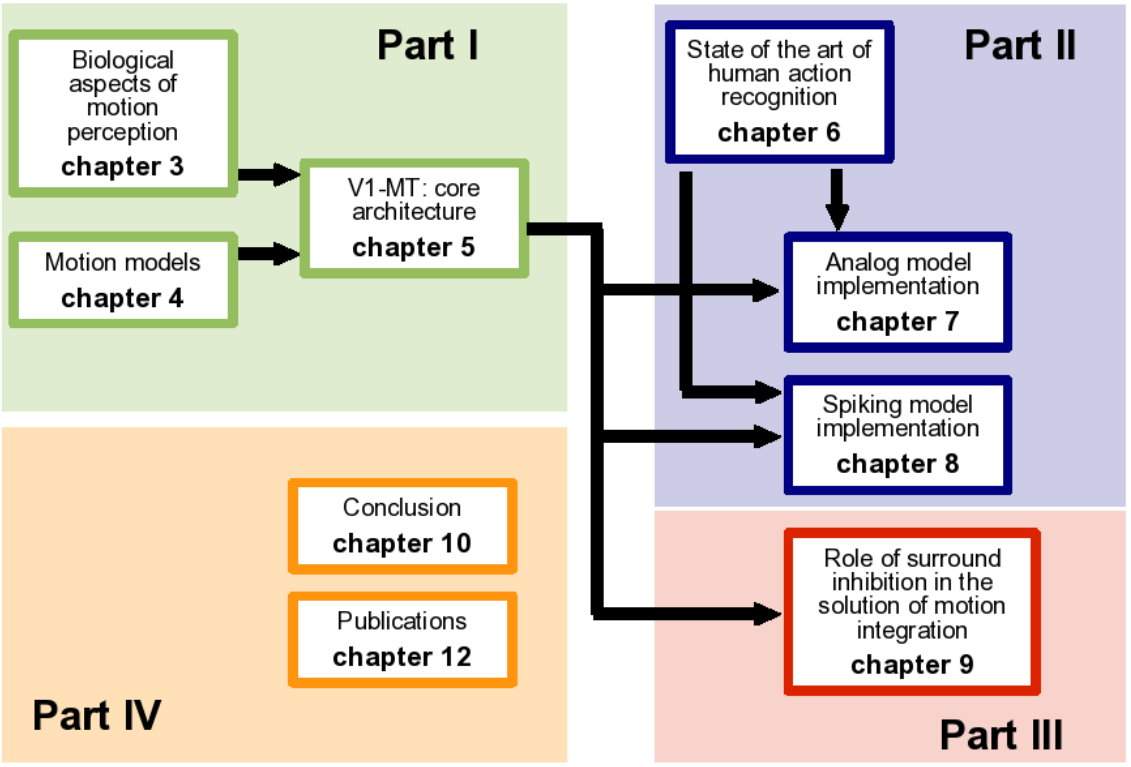


Figure 1.2: Organization of this thesis and relationship between chapters.

The main contributions are:

1. Proposition of a bio-inspired feedforward V1-MT core architecture.

2. Frequency-based analysis of the V1 motion detectors. This analysis shows the relationship between different parameters and the spatiotemporal frequency tuning.

3. The implementation of an analog V1-MT feedforward architecture to be applied to human action recognition. Together with the implementation of a classification method to analyze MT analog output. Also, the importance of MT surround diversity in human action recognition performance.

4. The implementation of a spiking V1-MT feedforward architecture to be applied to human action recognition. Considering two characteristics of the neural code: mean firing rate of each neuron and synchrony between pairs of neurons, two motion maps are defined as a representation of the input motion information: *mean motion map* and *synchrony motion map*. We show that these two motion maps can successfully perform human action recognition.

5. A simple mechanism to explain the shifting on the preferred-direction of MT neurons as a motion integration solution.

1.3 DETAILED PLAN

Part I: Motion Perception in Mammals

Chapter 3 gives the principles to understand how the visual motion information is processed in the mammal brain. This chapter describes the state of the art of the neurophysiological studies regarding motion perception in mammals, focusing on V1, MT and MST. This biological background will be a source of inspiration to develop bio-inspired models for motion processing.

The state of the art of motion detection and bio-inspired models for motion processing are described in **Chapter 4**. We review classical techniques to detect motion in input video sequences, both in computer vision and in computational neuroscience. In particular, we focused on those methods that inspired the development of models therein.

Considering the information reviewed in Chapters 3 and 4, **Chapter 5** describes the feedforward V1-MT core architecture proposed in this thesis for motion processing.

Part II: Human Action Recognition

The state of the art of the human action recognition is described in **Chapter 6**, covering the methods developed in the computer vision and computational neuroscience communities.

In order to deal with the human action recognition problem, we extended the feedforward V1-MT core architecture described in Chapter 5 (as it is shown in Figure 1.3) proposing as contribution two different implementations:

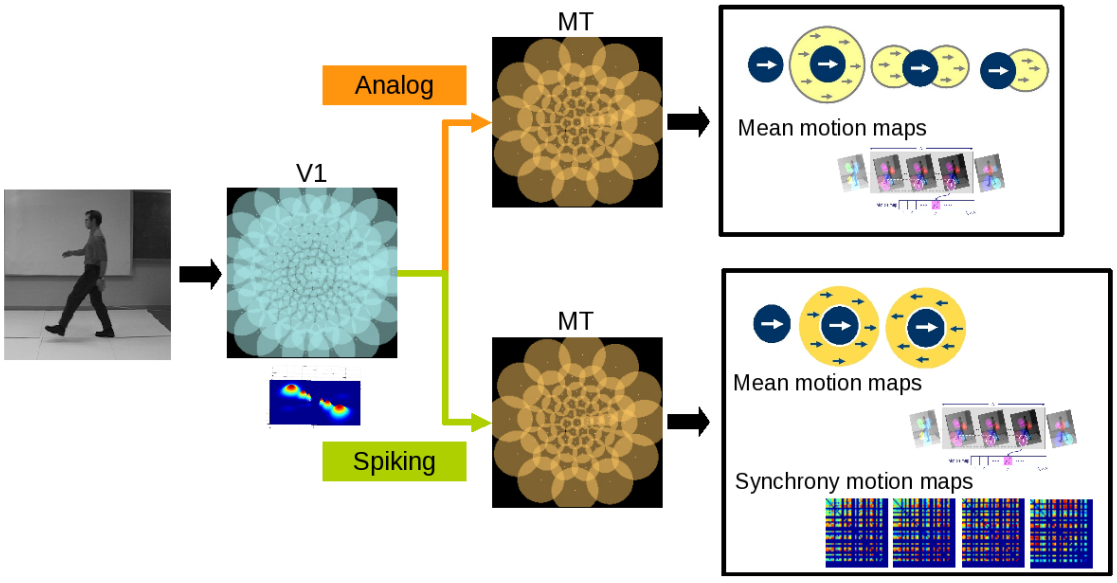


Figure 1.3: Towards human action recognition problem. Two implementations of the motion stream are investigated starting from the feedforward V1-MT core architecture defined in Chapter 5.

- **Chapter 7: Analog implementation.** The output of the energy-motion detectors is passed through a nonlinear function in order to estimate the mean firing rate of a V1 neuron. The output of V1 neurons feed a MT neuron which is modeled by a conductance-based neuron model. Depending on the spatial location of V1 neurons inside MT receptive field, V1 neurons can contribute as an excitatory or inhibitory conductance. The values of the membrane potential of MT neurons are used to define a feature vector (*mean motion map*) representing the motion information of the input sequence. These motion maps are then used to perform recognition.

- **Chapter 8: Spiking implementation.** The output of the energy-motion detectors feed a leak-integrate-and-fire (LIF) V1 neuron as an external input current. When the membrane potential of the V1 neurons reaches a threshold, a *spike* is generated. The spike trains obtained in V1 neurons feed spiking MT neuron layer through input conductance. Depending on the spatial location of V1 neurons inside MT receptive field, V1 neurons can contribute as an excitatory or inhibitory conductance. The spike trains generated by MT neurons are used to build two different motion maps representing the input motion information: *mean motion map* and *synchrony motion map*. The performance of both motion maps was evaluated in the human action recognition task.

Part III: Motion Integration

Chapter 9 explores the role of V1 surround suppression in the motion integration problematic to solve the aperture problem, specifically, how the V1 surround suppression can affect the preferred direction of MT neurons. The architecture implemented also derives from the feedforward V1-MT core architecture presented in Chapter 5, and it is shown in Figure 1.4. We show some simulations performed for classical psychophysics stimuli, such as barberpoles and plaids.

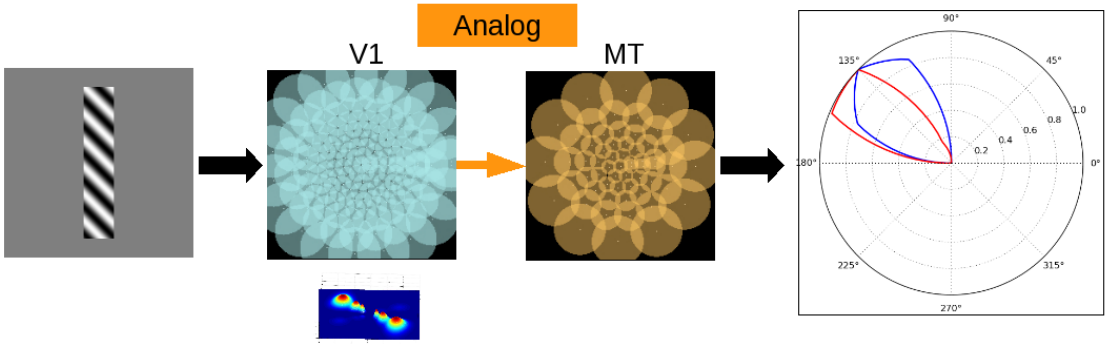


Figure 1.4: Study of the effect of V1 surround suppression in the solution of the aperture problem.

Part IV: Conclusion

Chapter 10 shows the conclusion, discussion and perspectives of the results obtained in this thesis and their comparison with the existing bibliography. Finally, **Chapter 12** enumerates the publications of the author arising from this thesis.

CHAPTER 2

INTRODUCTION (FRANÇAIS)

"What does it mean to see? The plain man's answer would be, to know what is where by looking. In other words, vision is the process of discovering from images what is present in the world, and where it is."
– David Marr

2.1 QUE SIGNIFIE VISION?

52%

52% est le pourcentage des aires corticales chez le singe macaque dédiées à la vision. Nous pouvons comparer ce pourcentage avec les autres modalités, somatosensorielle: 10%, moteur: 8%, auditif: 3%, olfactif: 1%. Donc, pour quelle raison le traitement de l'information visuelle est-il si exigeant?

David Marr, un pionnier de l'étude de la vision, a remarqué que la vision n'est pas seulement l'action de voir, mais l'action de traiter ce qui est présent dans le monde et où le système évolue. Selon lui, l'étude de la vision ne doit pas seulement inclure l'étude des proprietés isolées, mais aussi la nature des représentations internes grâce auxquelles nous sommes capable d'acquérir cette information et la rendre disponible pour nos décisions et actions.

La compréhension de la vision nécessite de comprendre le système visuel, tâche qui est un des projets les plus ambitieux des derniers 50 ans. Mesurer l'activité du cerveau, connecter des éléctrodes, classifier les différentes aires corticales selon leur fonctionnalité, etc, tous ces efforts vont dans la même direction, à savoir mieux comprendre les traitements corticaux, du niveau fonctionnel jusqu'au niveau neuronal.

Grâce à de nombreuses études neurophysiologiques, la connexion entre différentes aires corticales a été établie en révélant la grande complexité du système visuel, comme illustré à la Figure 2.1. Par exemple, parmi les différentes aires liées à la vision, il a été aussi possible d'identifier celles qui sont reliées au traitement du mouvement, les plus importantes etant: V1, MT et MST.

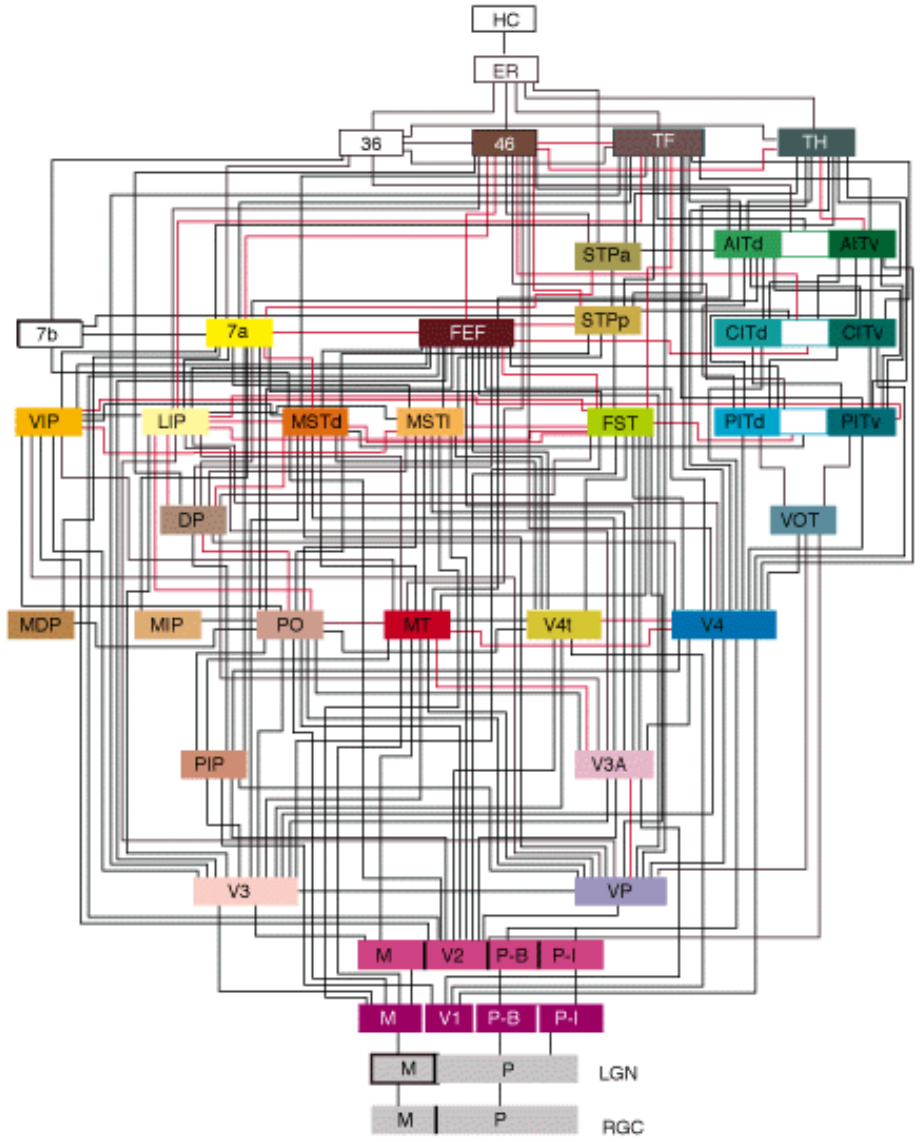


Figure 2.1: Diagramme des connexions du système visuel chez le singe macaque (extraite de Felleman and Van Essen (1991)).

Le Mouvement

Parmi les mécanismes reliés à la vision, l'analyse du mouvement est l'un des plus importants. La detection du mouvement est essentielle pour le survi de toutes les créatures. Les objets qui se déplacent pourraient être de probables prédateurs, source de nourriture, ou compagnon. En fait, l'effet de produire une activation en l'absence de mouvement est une proprieté des yeux qui apparaît vers la fin la chaîne évolutionnaire. Mais le mouvement est aussi essentiel pour de nombreuses autres tâches. Par exemple, considérons le cas d'une patiente ayant sa région cérébrale extrastiée endommagée, région que l'on pense reliée au traitement du mouvement. Cette patiente est incapable d'apprécier le mouvement des objets, mais elle a aussi des difficultés pour verser du thé dans une tasse parce que le liquide lui semble "gelé", ou

pour suivre un dialogue parce qu'elle n'est pas capable de suivre les mouvements des lèvres de la personne qui parle. Ceci illustre donc par exemple le rôle du mouvement dans nos activités quotidiennes, et aussi dans nos interactions sociales.

Le mouvement est aussi essentiel pour de nombreuses applications artificielles impliquant des données visuelles. En fait, tout système réel a besoin d'une étape d'estimation du mouvement. Par exemple, nous pouvons citer la robotique, la navigation autonome, prévision en météorologie, la restauration de vidéos, la récuperation de contenus, la vidéo surveillance, l'analyse d'une foule ou la reconnaissance d'actions. Donc, nous avons besoin d'avoir des algorithmes robustes et précis pour estimer le mouvement d'une scène réelle, et ces besoins justifient la littérature étendue sur le domaine produite par la communauté en vision par ordinateur. Cette problématique est encore ouverte la plupart des modèles reposant sur des hypothèses et contraintes concernant la scène à analyser. Aussi, nous pouvons nous demander si de nouvelles idées provenant de la biologie peuvent améliorer les performances des systèmes actuels.

Objet de cette thèse et méthodologie

Pour chercher à comprendre les fonctionalités du traitement du mouvement, l'objet de cette thèse est de proposer un système séquentiel pour l'analyse du mouvement. Selon notre définition, le terme bio-inspiré est associé aux modèles dans lequels soit l'architecture du cerveau a inspiré celle du système, soit des fonctionalités ou opérations réalisées par des cellules réelles sont implementées, par exemple une implémentation analogique ou événnementielle ("spikante"). L'implémentation de modèles bio-inspirés peut nous donner une meilleure compréhension des proprietés du système visuel. Cependant seule une petite partie des modèles existant s'appliquent à des séquences réelles.

L'objet de cette thèse est le traitement du mouvement. Nous étudions les mécanismes liés avec cette tâche chez le mammifère et nous proposons des modèles sequentiels modélisant les aires corticales V1-MT, pour affronter deux problèmes classiques de l'analyse du mouvement: la reconnaissance d'actions et l'intégration du mouvement.

La reconnaissance d'actions

La reconnaissance d'actions est la tâche de reconnaître une action enregistrée dans une séquences d'images, donc d'assigner à chaque séquence une étiquette dénotant l'action, ex: marcher, courir, etc...

Initialement, l'analyse du mouvement chez l'humain a été étudié par le psychologue Gunne Johansson en 1973 (Johansson (1973)). Il a demontré comment très peu d'information est requis pour percevoir le mouvement, chez l'humain, et chez les animaux. Il a placé des points lumineux sur les bras et les jambes d'une personne.

Quand la personne bouge, le mouvement des points lumineux est identifié comme un mouvement humaine, dont le genre peut même être identifié. Ce type de stimulus, connu sur le nom de *biological motion* ou *point-light stimulus* est un pattern de mouvement complexe et intéressant pour plusieurs raisons: 1) il montre un lien entre la perception du mouvement et la forme (Pucel and Perret (2003); Michels et al. (2005); Hirai and Hiraki (2006)), 2) il évite les distracteurs comme le fond, les vêtements, etc.

Mais le monde réel n'est pas un stimulus de type *point-light*: Notre système visuel reçoit comme entrée un flux continu d'images qui doit être interpreté dans sa globalité avant de procéder à la reconnaissance des *patterns* de la scène. Puis, selon cette reconnaissance, nous interagissons avec l'environement. La plus grande partie de notre vie sociale et de nos interactions avec l'environnement vient de la reconnaissance des membres de notre entourage, particulièrement des actions réalisées par eux. En tant qu'humain, nous pouvons reconnaître facilement si une personne se rapproche en marchant, en courant, ou bien si cette personne nosu salue avec une ou ses deux mains.

L'automatisation de cette tâche de reconnaissance a plusieurs applications dans différents domaines, incluant la vidéo surveillance et l'interaction homme-machine. Cette problématique est traitée par la communauté de vision par ordinateur, où les méthodes développées n'ont pas vocation à s'inspirer de modèles biologiques..

Actuellement, entre les méthodes proposées dans les communautés de vision par ordinateur et dans les neurosciences computationnelles, aucune n'est capable de reconnaître des actions dans n'importe quelles conditions et scènes. Cette limitation motive le dévelopement de nouvelles approches qui suivent de nouvelles pistes. Nous nous demandons ici si le dévelopement de méthodologies inspirées par la biologie peut nous montrer ces nouvelles pistes.

Dans le cadre de cette thèse, nous avons developpé deux architectures différentes pour accomplir la reconnaissance d'actions: Nous avons exploré comment les actions peuvent être représentées à partir de la sortie analogique ou évènementielle des neurones de MT. Par exemple, comment l'implémentation de différentes interactions entre centre et périphérie dans les neurones de MT a une incidence sur la performance de la reconnaissance d'actions?

Intégration du mouvement

Les mécanismes d'intégration du mouvement offrent une solution au problème d'ouverture. Le problème d'ouverture est un problème classique et plusieurs modèles ont été proposés comme solution. Le problème d'ouverture a attiré l'attention des scientifiques qui ont aussi observé que le mouvement réel d'objets est relatif, et que la seule façon de le mesurer est à partir de l'information relative des objets autour.

Plusieurs mécanismes ont été proposés pour chercher à comprendre comment le système visuel chez le mammifère calcule la vraie direction du mouvement des objets. Parmi les mécanismes proposés, nous pouvons citer: l'information de détection

de contours provenant de V2, la diffusion spatiale des indices de mouvement non-ambigüs, la rétroaction d'autres couches neuronales, l'inhibition périphérique, des mécanismes de winner-take-all, etc.

À partir de l'inspiration donnée par les résultats rapportés par Pack et al. (2004), dans cette thèse, nous nous sommes demandés comment la suppression périphérique est liée à la solution du problème d'ouverture. La suppression du poutour dans les cellules de V1 a été implementée dans l'architecture séquentielle de V1-MT proposée dans le cadre de cette thèse, et nous avons exploré l'effet de cette suppression sur la direction préferée des neurones de MT. Ce mécanisme a été précédemment reporté comme une possible solution au problème d'ouverture (Pack et al. (2004); Huang et al. (2007); Tlapale et al. (2008)).

2.2 Organisation et principales contributions

Cette thèse est organisée en quatre parties (Figure 2.2):

- Première Partie: *La perception du mouvement chez le mammifère.* Dans cette première partie nous résumons les études concernant le traitement de l'information visuelle du mouvement chez le mammifère et comment les mécanismes de vision ont été modélisés de manière classique en terme de détection et traitement. Nous présentons aussi l'architecture séquentielle générale pour V1 et MT proposée et dévelopée dans cette thèse.

- Deuxième Partie: *La reconnaissance d'actions.* Nous montrons dans cette partie l'état de l'art de cette problématique et proposons aussi deux différentes approches, créées à partir de l'architecture séquentielle génerale décrite dans la première partie.

- Troisième Partie: Cette partie montre comment l'implémentation de la suppression périphérique dans les neurones de V1 peut être impliquée dans la résolution du problème d'ouverture.

- Finalement, la Quatrième Partie se groupe la conclusion, les perspectives et les publications associées à ce travail.

Les principales contributions de cette thèse sont:

1. La proposition d'une architecture séquentielle bio-inspiré, générale, modélisant les aires corticales V1 et MT.

2. Une analyse fréquentielle des détecteurs de mouvement sur V1. Cette analyse montre la relation entre les différents paramètres et le réglage fréquentiel spatio-temporel.

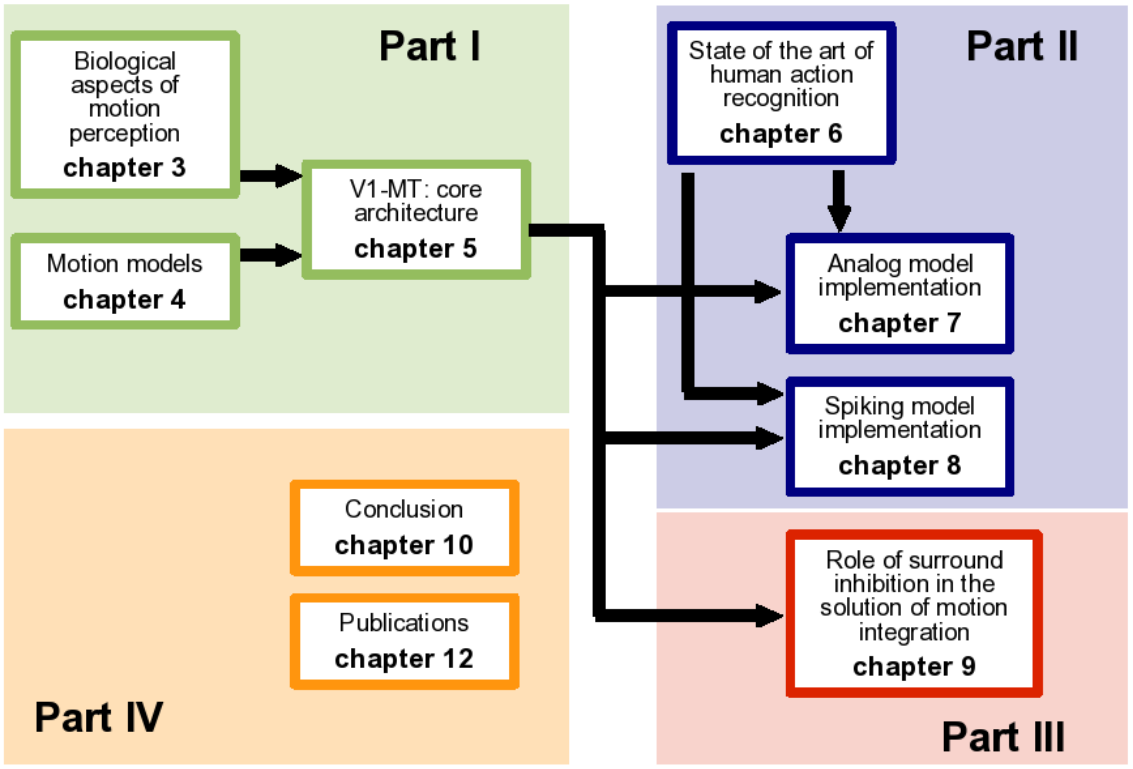


Figure 2.2: Organisation de cette thèse et relation entre chapitres.

3. L'implémentation d'une architecture séquentielle analogue modélisant V1 et MT pour être appliquée à une application réelle comme la reconnaissance d'actions. Nous présentons une méthodologie pour le classement à partir de l'analyse de la sortie analogique des neurones de MT. Nous mettons en évidence l'importance de la diversité de périphéries dans les neurones de MT pour la performance de la reconnaissance d'actions.

4. L'implémentation d'une architecture séquentielle évènementielle modélisant V1 et MT pour être appliquée à une application réelle comme la reconnaissance d'actions. Nous avons considéré deux caractéristiques du codage neuronal: le taux moyen de décharge de chaque neurone et la synchronie entre paires de neurones. Deux *motion maps* sont donc définis, donnant une répresentation de l'information du mouvement contenue dans le stimulus: *mean motion map* and *synchony motion map*. Nous montrons que ces deux *motion maps* permettent de réaliser avec succès la reconnaissance d'actions.

5. L'implémentation d'un mécanisme simple pour expliquer le changement de la direction préferée d'un neurone de MT, et aussi, pour donner une solution à l'intégration du mouvement.

2.3 PLAN DETAILLÉ

Première Partie: La Perception du Mouvement chez le Mammifère

Le **Chapitre 3** donne les principes de la compréhension du traitement d'information visuelle reliée au mouvement chez le mammifère. Ce chapitre décrit aussi l'état de l'art des études neurophysiologiques par rapport à la perception visuelle chez le mammifère, en mettant l'accent sur V1, MT et MST. Ce cadre biologique sera la source d'inspiration du dévelopement des modèles bio-inspirés pour le traitement du mouvement.

L'état de l'art de la détection de mouvement et de modèles bio-inspirées pour le traitement du mouvement est décrit dans le **Chapitre 4**. Nous avons récapitulé les techniques classiques en vision par ordinateur et en neurosciences computationelles pour la detection du mouvement. En particulier, nous nous sommes focalisés sur les méthodes qui ont inspiré le dévelopement de cette thèse.

A partir de l'information récapitulée dans les Chapitres 3 et 4, le **Chapitre 5** décrit l'architecture séquentielle générale modélisant les aires corticales V1 et MT proposée dans cette thèse pour le traitement du mouvement.

Deuxième Partie: La Reconnaissance d'Actions

L'état de l'art de la reconnaissance d'actions est décrit dans le **Chapitre 6**. Cet état de l'art couvre les méthodes developées en vision par ordinateur et en neurosciences computationnelles.

Afin de traiter le problème de la reconnaissance d'actions, nous avons étendu l'architecture séquentielle générale proposée pour V1 et MT décrite dans le Chapitre 5 (Figure 2.3), et proposons comme contribution originale deux implementations.

- **Chapitre 7: Implementation analogique.** La sortie des détecteurs de mouvement basés sur l'énergie traverse une fonction non-linéaire pour obtenir une estimation du taux de décharge des neurones de V1. La sortie des neurones de V1 alimente un neurone de MT modelisé par un modèle neuronal à conductance. Selon la localisation spatiale des neurones de V1 dans le champ récepteur du neurone de MT, le neurone de V1 peut apporter une conductance excitatrice ou inhibitrice. Les valeurs des potentiels de membrane des neurones de MT sont utilisés pour définir un vecteur de mouvement moyen (*mean motion map*) qui répresente l'information de mouvement contenue dans la séquence d'entrée. Ces *mean motion maps* sont utilisés pour la reconnaissance.

- **Chapitre 8: Implémentation évènementielle.** La sortie de détecteurs de mouvement basés sur l'énergie alimente un neurone de V1 par un courant externe. Les neurones de V1 sont modelisés comme des neurones intègre-et-tire

à fuite (LIF). Quand le potentiel de membrane d'un neurone de V1 atteint un seuil, une impulsion *spike* est émise. Les trains de *spikes* obtenus à partir des neurones de V1 alimentent un neurone de MT, évènementiel aussi, au travers d'une conductance d'entrée. Selon la localisation spatialle des neurones de V1 dans le champ récepteur d'un neurone de MT, le neurone de V1 peut apporter une conductance excitatrice ou inhibitrice. Les trains de *spikes* produits par les neurones de MT sont utilisés pour définir deux cartes de mouvement *motion maps* différents: carte de mouvement moyen (*mean motion map*) et carte de synchronie (*synchrony motion map*). La contribution de chaque carte est évaluée dans la tâche de reconnaissance d'actions.

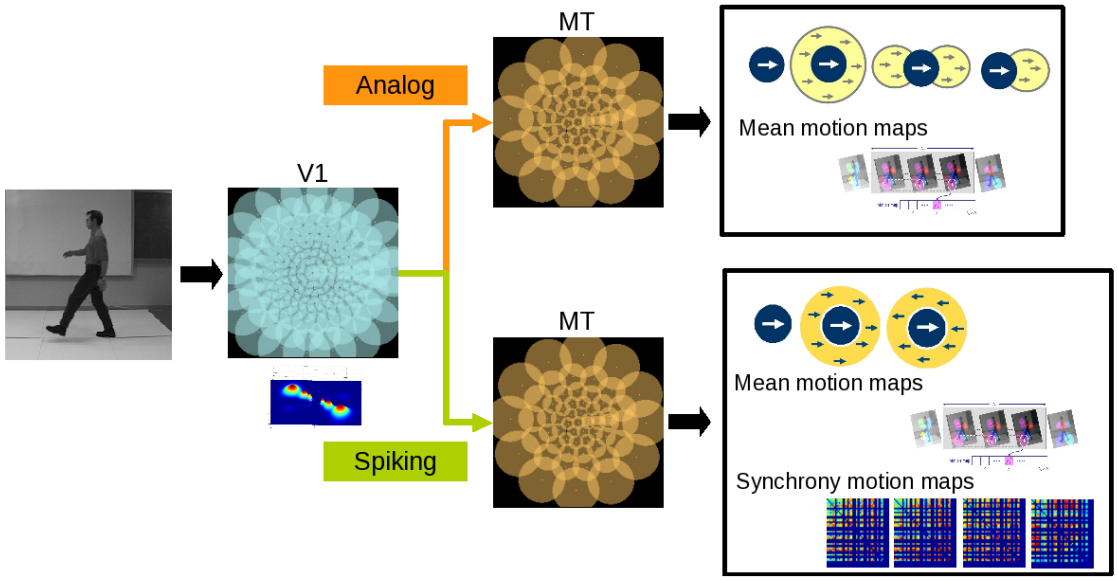


Figure 2.3: Reconnaissance d'actions. Deux implémentations pour la modélisation du flux de mouvement sont examinées. Ces deux implémentations sont faites à partir de l'architecture générale de V1 et MT, proposée dans le Chapitre 5.

Troisième Partie: L'Intégration du mouvement

Le **Chapitre 9** explore le rôle de la suppression périphérique des neurones de V1 sur l'intégration du mouvement, ainsi que la résolution du problème d'ouverture. De manière plus précise, nous nous demandons comment la suppression périphérique des neurones de V1 peut affecter la direction préferée des neurones de MT. L'architecture implémentée provient aussi de l'architecture générale de V1 et MT présenté dans le Chapitre 5, et elle est montrée à la Figure 2.4. Nous donnons les résultats des simulations faites pour plusieurs stimuli classiques en psychophysique, comme les barberpoles et les plaids.

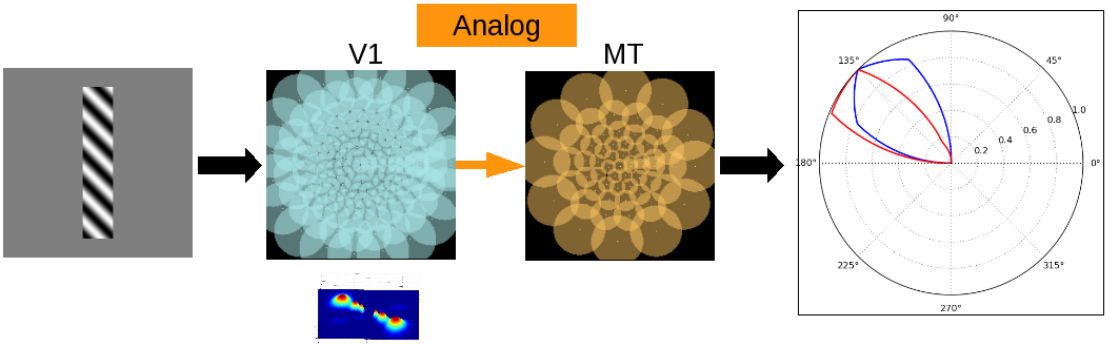


Figure 2.4: Étude de l'effet de la suppression périphérique des neurones de V1 sur la solution du problème d'ouverture.

Quatrième Partie: Conclusion

Le **Chapitre 10** presente une large discussion autout les résultats obtenus dans ce travail ainsi que nos perspectives.

Le **Chapitre 12** indique les publications de l'auteur de cette thèse.

Part I

Motion Perception in Mammals

CHAPTER 3

BIOLOGICAL ASPECTS OF MOTION PERCEPTION

Contents

OVERVIEW

How motion information is processed in the visual system of mammals? The perception of motion in the visual system involves many brain areas, which have been widely studied along years. The main brain areas dedicated to motion analysis in mammals are V1, MT/V5 and MST. V1 is the most studied area of the visual system of mammals and we could say that it is the gate from visual sensory system to the brain. V1 processes not only motion but also other visual characteristics, such as shape, color and texture. MT and MST are apparently only motion sensitive and MT can be easily called as "the motion brain area".

In both areas V1 and MT, neurophysiological studies have shown a high influence of what is seen in the *surround* of their respective receptive fields. What is seen in the *surround* clearly modulates the responses of V1 and MT neurons and also might be related with the visual perception.

In this chapter we revisit the state of the art concerning the motion related areas V1, MT and MST, describing what do we know about them in the motion processing context and also how the *surround* information influences the V1 and MT cells' response. The concepts here described will be used in the next chapters of this thesis for our V1-MT model implementation.

Keywords: V1, MT, MST, motion perception, motion processing, early vision, classical receptive field (CRF), center-surround interactions, surround suppression.

Organization of this chapter:

This chapter is organized as follows. Section 3.1 describes the state of the art of neurophysiological studies of V1 regarding motion processing and also the different surround interactions. Analogously, Section 3.2 shows the state of the art of MT neurons together with their surround interactions. Finally, Section 3.3 describes a brief state of the art of MST visual area.

3.1 V1: EARLY MOTION ANALYSIS

Located in the occipital lobe, the primary visual cortex V1 (also known as area 17 or striate cortex) is far the most studied visual area of primates. It is ambitious to propose a general state of the art which renders the hundreds of works studying all its different aspects. So, considering only the interesting aspects for this thesis, which is motion processing, this brief review will include the main functional properties of V1 regarding motion processing and center-surround interactions.

V1 is the first visual area processing the visual information coming from the retina and passing through the lateral geniculate nucleus (LGN). It is divided into six different functional layers, from layer 1 up to layer 6. Inputs coming from the LGN are received by layer 4. The magnocellular input from the LGN is received by Layer 4Cα, while the parvocellular inputs are received by layer 4Cβ. This fact inspired the idea of two different cortical system to process the visual information (Ungerleider and Mishkin (1982); Goodale and Milner (1992); Milner and Goodale (2008)): one specialized in motion perception located at the parietal lobe (*dorsal stream* concerning V1, V2, MT, MST, LIP, VIP and PP), and one specialized in form perception located at temporal lobe (*ventral stream* concerning V1, V2, V4, PIT or TEO, AIT or TE). This simplified structure is pedagogically convenient but it has been widely criticized (see e.g., Van Essen and Gallant (1994); Milner and Goodale (2008)).

Differing from the retina and LGNs whose receptive fields have round shapes, the receptive fields of V1 simple cells have elongated shapes. They are supposed to be built from thalamocortical afferents (Hubel and Wiesel (1962); Chapman et al. (1991); Alonso et al. (2001); De Valois et al. (2000)) and intracortical process of amplification and/or inhibition (Ferstner and Miller, 2000). Two main ideas have been proposed to understand how the elongated receptive fields are formed:

1. The cortical receptive fields are simply cortical manifestations of the centers and surrounds of receptive fields of retinal neurons.

2. The cortical receptive fields are a combination of inputs from separated sets of receptive fields whose centers have opposite polarity.

The work done by Kara and Reid (2003) have shown that cortical subfields are formed from the centers of retinal receptive fields (see Figure 3.1).

Motion is processed by direction-selectivity neurons, but how are the inputs of these neurons? Classically, the M and P pathways coming from the LGN, differed in space and time, are required to create direction selectivity. The P pathway has receptive field with sustained response, while the M pathway has receptive fields with a transient response and differs from the P pathway in temporal phase by about a quarter cycle. This characteristics suggest that combining these two streams could lead

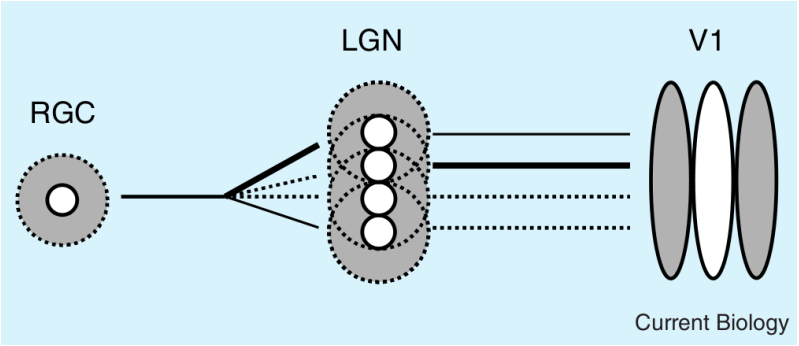


Figure 3.1: The pathway from an individual retinal output neuron (retinal ganglion cell (RGC)) to the cortex involves divergence in the LGN through multiple pathways of different strengths (indicated by the thickness of lines; dashed lines indicate unknown strength), which could re-converge in the cortex, specifically in V1 (Image taken from Derrington and Webb (2004), adapted from Kara and Reid (2003)).

direction selectivity. In fact, De Valois et al. (2000) stated that direction-selectivity of simple cells can be decomposed into a fast, temporally biphasic, spatially even symmetric component and a slower, temporally monophasic, spatially odd symmetric component. This decomposition can give an idea of how the direction-selectivity property is formed. Livingstone and Conway (2003) found similar receptive field maps than De Valois et al. (2000), but they suggest that the slow component comes from a delayed offset inhibition rather than the slow parvocellular pathway. The later work of Saul et al. (2005) hypothesizes that not only the M and P pathways by separated contribute to create the direction-selectivity property, but a combination of both.

3.1.1 Simple and complex cells

"Indeed, all that may distinguish many complex cells from simple cells might just be the strength of the inhibitory signals that mask inherently nonlinear summations..."
–Carandini et al. (2005)

The seminal work of Hubel and Wiesel (1960, 1962) studied and classified V1 neurons into two groups: *simple cells* and *complex cells*. These cells react to stimuli placed at a certain region of the visual field (around 1° of diameter) and they are sensible to the luminance or contrast of the stimuli. Hubel and Wiesel (1962) discovered that these cells' receptive fields are made up of elongated and antagonistic zones (see Figure 3.2). Depending of the sign of each zone, they sum the light falling on it to generate excitatory or inhibitory contributions to the cell activation.

According to Hubel and Wiesel (1962), a V1 *simple cell* is a cell who accomplishes these four conditions:

1. It can be divided into distinct excitatory and inhibitory regions.

2. The information inside the excitatory and inhibitory regions is summed.

3. There is antagonism between excitatory and inhibitory regions.

4. It is possible to predict responses to stationary or moving spots of various shapes from a map of the excitatory and inhibitory areas.

If a cell failed in at least one of these conditions, it is straight classified as a *complex cell*. As a corollary of this definition, we can say that in simple cells excitatory and inhibitory responses are spatially separated and mutually antagonistic. By the contrary, in complex cells the excitatory and inhibitory regions overlaps and are not mutually antagonistic (see Figure 3.2).

Since Hubel and Wiesel (1962), additional properties of simple and complex cells have been found refining their definitions. For example, Conway and Livingstone (2003) stated that simple cells space-time maps have an overall slant, while complex cells showed space-time maps not clearly slanted. Complex cells have higher firing rate than simple cells. Complex cell response is more transient than the simple cell response.

Although complex cells are also oriented selective, it is evident that this selectivity cannot be deduced from conventional receptive fields maps (Conway and Livingstone (2003); Deangelis and Akiyuki (2004)). In the complex cell receptive field maps bright and dark responsive regions overlap almost completely in the spatiotemporal domain, and no distinct regions are visible (see Figure 3.2 C). Using a nonlinear map, built with 2-dimensional white noise, it is possible to predict the direction selectivity of V1 complex cells, whereas their linear maps do not (Deangelis and Akiyuki (2004)).

Studies in the literature suggest that complex cells are built by the nonlinear combination of subunits, e.g., simple cells (Hubel and Wiesel (1962); Movshon et al. (1978); Emerson et al. (1987); De Valois et al. (2000); Pack et al. (2006)).

Using reverse correlation, De Valois et al. (2000) characterized the shape of the receptive fields of directionally and non-directionally oriented V1 simple cells (see Figure 3.4). Different maps for V1 simple and complex cells were also measured by Conway and Livingstone (2003) and Pack et al. (2006). The shape of the receptive fields of V1 simple cells can be typically modeled by Gabor functions. Specifically, Ringach (2002) showed that Gabor can provide a representation for the receptive fields of V1 simple cells measured in monkeys (see Figure 3.3). Pack et al. (2006) found similarities between the receptive field maps of V1 complex cells and MT suggesting that MT receptive fields are primarily built by summing the outputs of V1 complex cells sharing a common preferred direction.

Regarding the spatiotemporal frequency and speed tuning of V1 simple and complex cells, Priebe et al. (2006) measured the response of direction-selectivity V1 simple and complex cells of anesthetized, paralyzed macaque monkeys. The direction-selectivity of V1 simple cells showed separable tuning for spatial and temporal frequencies, while V1 direction-selectivity complex cells showed the same speed tuning

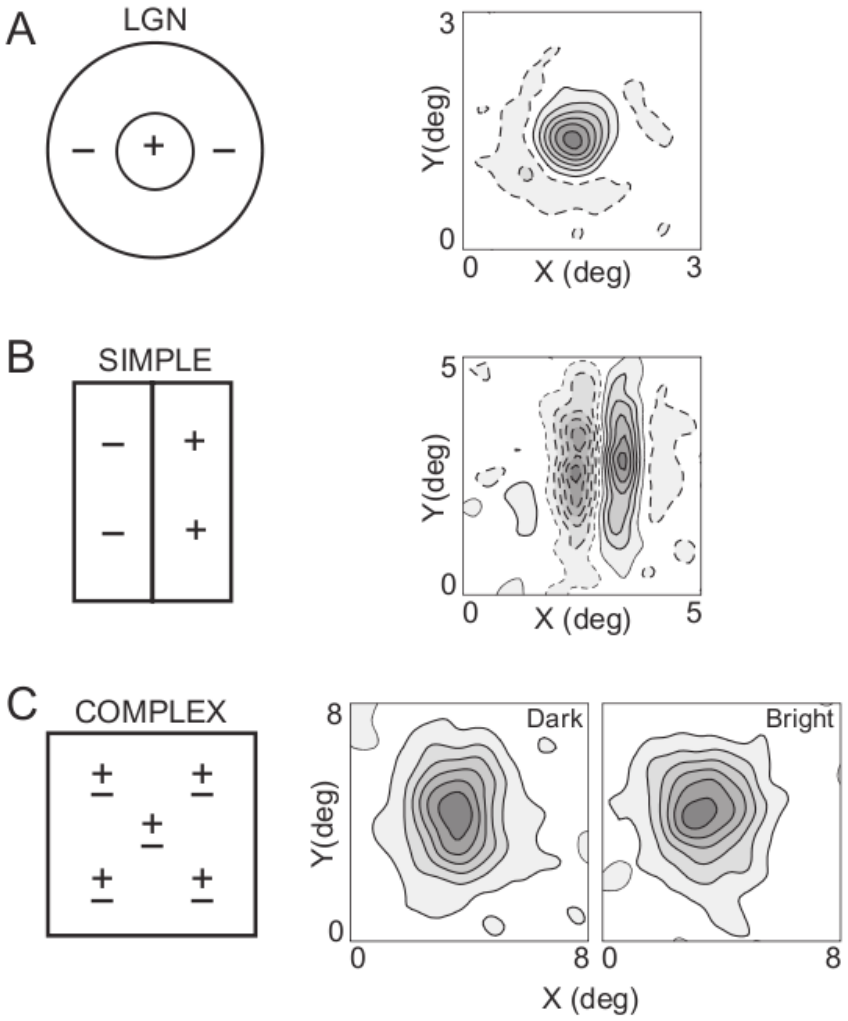


Figure 3.2: Receptive fields structure of LGN, simple and complex V1 neurons in cats. **A**, Schematic and empirical receptive field of a **LGN neuron** from cat. It has a central ON region (+) surrounded by an OFF region (-). Solid and dashed regions represent regions in the visual space where the cell responds to bright or dark spots, respectively. **B**, Schematic and empirical receptive field of a **V1 simple cell**, which its receptive field consists of alternating elongated subregions that are responsive to bright (+) or dark (-) visual stimuli. **C**, Schematic and empirical receptive field of a **V1 complex cell**. This type of cell responds to both bright and dark stimuli anywhere inside its receptive field (Image taken from Deangelis and Akiyuki (2004)).

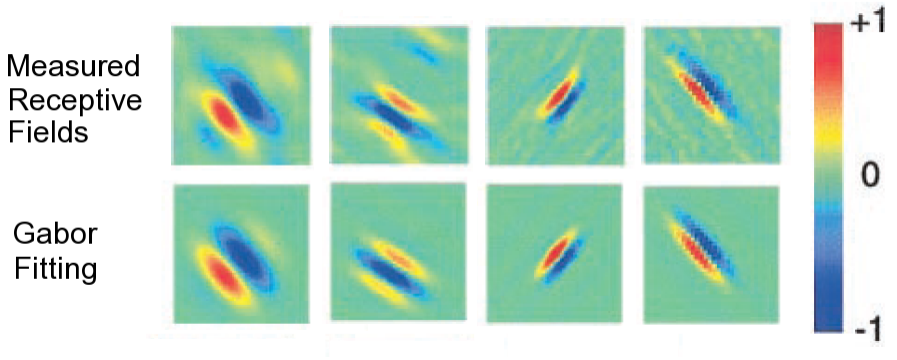


Figure 3.3: Two-dimensional Gabor functions fitting V1 simple cell data. First row shows examples of receptive fields measured for V1 simple cells. Second row shows the best Gabor fit in the least square sense, showing that Gabor functions can represent the shape of V1 receptive fields (Image adapted from Ringach (2002)).

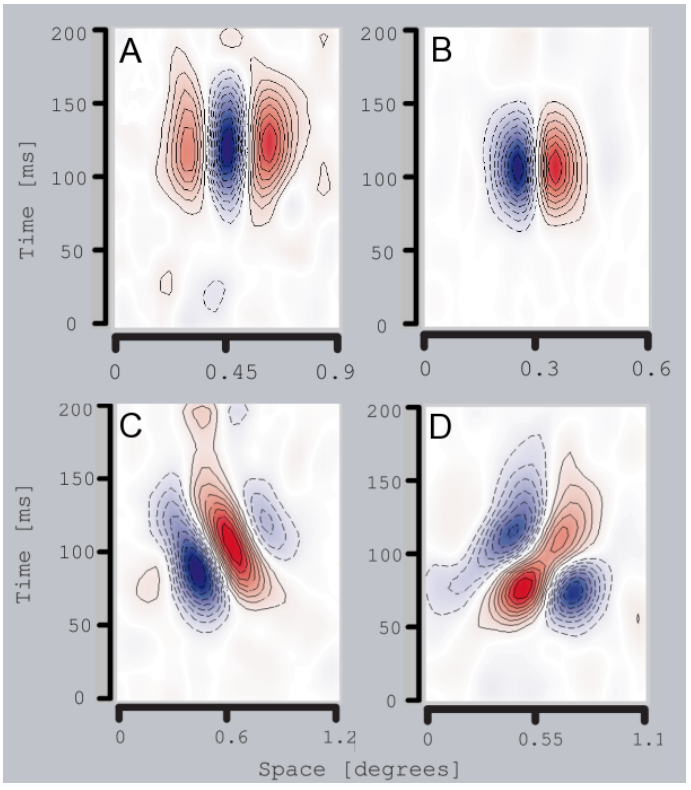


Figure 3.4: Examples of the spatiotemporal receptive fields of two non-directionally-selective V1 simple cells (**A**, **B**), and two directionally-selective V1 simple cells (**C**, **D**) (image adapted from De Valois et al. (2000)).

properties found in area MT (see Priebe et al. (2003)). Concerning V1 complex cells, the cells measured can be classified as: $\sim$25% with separable responses to spatial and temporal frequencies, $\sim$25% speed-tuned neurons with a preferred speed that does not depends on the spatial frequency, $\sim$50% between these two extremes.

An important nonlinear effect seen in V1 neurons is the cross-orientation suppression (COS). The V1 neuron response to an optimally stimulus is inhibited if an orthogonal stimulus is superimposed (transparency). This orthogonal stimulus does not elicit a neural response if it is presented alone. Studies in cats (Morrone et al. (1982)) showed this effect using two superimposed drifting gratings. This type of inhibition is typically stronger in simple than complex cells. They also showed that this effect is not only cross-oriented, because not only the orthogonal orientation but all orientations outside the cell's tuning band have a comparable inhibitory effect. This last effect suggests that this type of inhibition comes from not only a single cell but a population of cells. Regarding possible origins of the COS, Li et al. (2006) stated that nonlinearities in the LGNs, such as, spike rectification and contrast saturation plus a spike output nonlinearity in the visual cortex could explain the COS effect.

3.1.2 Center-surround interactions

Receptive fields of V1 neurons can be decomposed into a classical receptive field (CRF), in which stimuli directly elicit the discharge of the neuron, and a large surrounding area beyond the CRF, in which stimuli elicit no response by their own but they can profoundly modulate the CRF-driven response, normally suppressing it (Jones et al. (2001)).

The majority of V1 simple and complex cells have surround suppression. In anesthetized monkeys, Jones et al. (2001) found that 94% of measured cells exhibited surround suppression with a mean suppression of 67%, and 43% of cells exhibited a suppression greater than 70%. They also showed that the surround modulation is not always suppressive, and that the 78% of their studied cells were sensitive to the direction of motion of the surround. Within these cells two groups were detected:

- A direction-contrast-dependent group (41%) where the suppression was of 70% if the direction of motion of the surround was iso-oriented compared to the CRF, and of 22% if the direction of motion of the surround was reversed compared to the CRF.

- A direction-contrast-driven facilitation group (37%) where a suppression of 28% was detected if the direction of motion of the surround was iso-oriented to the CRF. If the motion of the surround was reversed compared to the CRF, a facilitation of 74% was detected.

Apparently, there is no laminar variation in the cells showing surround suppression (Walker et al. (1999); Jones et al. (2001)). Also, the surround suppression be-

tween simple and complex V1 cells is not significantly different (Sceniak et al. (2001); Bair et al. (2003); Webb et al. (2005)).

Following in part the schema presented by Series (2002) in her thesis and Series et al. (2003), we will next describe the state of the art of the surround suppression phenomenon in V1 neurons.

Size and geometry of surround

The size of the surround depends on species. In monkeys, Sceniak et al. (2001) measured the size of the surround zone as 2.2 times the size of the CRF. Similarly, Angelucci et al. (2002) found the average size of the surround zone to be 4 times the size of the CRF. In cats, Li and Li (1994) found that the 70% of measured cells presented a maximal suppression for surrounding zones of $\sim$2-5 times the size of the CRF, the resting 25% of cells exhibited maximal suppression for suppression zones up to 5 times the size of the CRF.

The surround suppression zone is far to be uniform. Walker et al. (1999) in anesthetized cats found that most of cells with surround suppression have spatially asymmetric surrounds. More in details, the study of Jones et al. (2001) in monkeys showed that only the 19% of the measured cells exhibited a uniform surround suppression. The remaining 81% of cells have either spatially asymmetric surround suppression (44%) or bilateral symmetric surround suppression (37%). Within the surround areas, the suppression is nearly equally distributed, which is contrary to the findings of Walker et al. (1999) in cats, where the suppression has a slight bias to occur at the end zones of the CRF.

Recently, Tanaka and Ohzawa (2009) studied detailed spatial structures of classical center and surround regions of V1 receptive fields. They found that center and surround regions are often both elongated parallel to each other, showing a wide range of orientations and widths.

Response latencies and origins

Bair et al. (2003) measured in monkeys that the latency of the suppression effect depends on the suppression strength and it varies systematically across cells. Strong suppression arrived on average $\sim$30ms earlier than weak suppression, and suppression sometimes arrives faster than the excitatory CRF responses. The delay of surround suppression with respect to the CRF excitation has been reported to a range from 15 to 60ms. They detected two different mechanisms: an early suppressive mechanism (prominent at lower CRF contrasts, spatiotemporally broadband and monocularly driven) and a late suppressive mechanism (CRF driven by high contrast stimuli, sharp spatiotemporal tuning and binocularly driven).

In cats, Walker et al. (1999) showed that the CRF has a short latency to response onset ($\sim$20ms), a peak response around the 50ms followed by a sharp decayment in

Table 3.1: Average spatial and temporal frequencies measured by Webb et al. (2005) for the CRF and the surround zone of a population of V1 neurons. The temporal frequency response of the surround is almost flat inside the measured range.

	CRF	Surround
Average spatial frequency	2.98[cycles/degree]	1.35[cycles/degree]
Average temporal frequency	4.99[Hz]	9.23[Hz]

the response. On the contrary, the surround exhibited a short latency (∼30ms), a peak suppression response near 60ms and then the suppression response is sustained and remains observable until 150ms.

The origin of surround suppression remains unknown. Most of the theories about its origin are mainly based in the latency studies mentioned in the previous paragraphs. Considering latency studies, Smith (2006) has proposed three different origins:

1. Long-range lateral or horizontal connections within V1 (Angelucci and Bullier (2002)).

2. Feedbacks from higher cortical areas due to the slow dynamics of surround suppression and the lack of strong dependence on cortical distance (Bair et al. (2003); Angelucci and Bullier (2003); Schwabe et al. (2006)).

3. Activation of surround suppression in the lateral geniculate nucleus (LGN) which leads to reduce excitatory drive to V1 (Webb et al. (2005)).

Spatiotemporal surround tuning

The spatiotemporal tuning of the suppression zones surrounding the CRF are similar to the cell's excitatory center tuning but broader (Sceniak et al. (2001); Webb et al. (2005)). In particular, Webb et al. (2005) found that the suppression is relatively indifferent to the orientation of the grating of the surround, particularly when the surround was driven by high temporal frequencies and when the CRF was stimulated by gratings at low contrasts. The average spatial and temporal frequencies preferred for the CRF and the surround in the population of cells measured by Webb et al. (2005) can be seen in Table 3.1. Levitt and Lund (1997) showed that the orientation selectivity of the surround is less sensitive when the CRF is driven by low contrast stimuli. In cats, Walker et al. (1999) found that the minimal suppression was obtained when the surround was orthogonal to the preferred orientation of the CRF.

See later: *In Chapter 9 we model V1 center-surround interactions and we show how this mechanism can be used to solve the aperture problem.* ■

Contrast dependency

Levitt and Lund (1997) and Sceniak et al. (1999) found that the effect of the surround in the cell activity depends on contrast, for instance, identical stimulus configuration could elicit facilitatory or suppressive center-surround interactions depending on the contrast of the central stimulus. Levitt and Lund (1997) also found that the particular surround stimulus that produced the larger effect remained invariant with center contrast.

Not only the type of the center-surround interaction changes with contrast. More recently, Sceniak et al. (2002) showed that cells with contrast-dependent changes in spatial summation also have a spatial frequency tuning depending on stimulus contrast. Sceniak et al. (2002) also showed that reduction of stimulus contrast causes significant sharpening of spatial frequency selectivity. Lately, Walker et al. (1999) and Webb et al. (2005) have shown that the surround suppression monotonically increases with increased surround contrast.

3.2 MT: THE MIDDLE TEMPORAL AREA

The middle temporal visual area (MT or V5) of the macaque monkey is an extrastiate visual area in which most cells are selective for the direction of motion stimulus (Movshon and Newsome (1996)).

3.2.1 Organization and connectivity

MT is retinotopically organized with an emphasis in the fovea, where the half of MT surface is dedicated to the processing of the central 15° of the visual field. The size of the receptive fields increases with the eccentricity and it also depends on contrast (Pack et al. (2005)). At a given eccentricity, the MT receptive fields are about 6 to 10 times larger than those in V1 (Churchland et al. (2005)). Figure 3.5 shows the sizes of V1 and MT receptive fields depending on the eccentricity.

MT has been always associated with the dorsal stream assuming most of its input coming from the LGN magnocellular (M) pathway. But anatomical and functional studies have proved that the early parallel visual pathways (magnocellular (M) and parvocellular (P)) converge significantly onto both dorsal and ventral cortical areas. In the case of MT, M and P pathways are present. P connections from LGN are only two synapses away and they probably come from V2. Still, M connections are the major ascending inputs which arrive through V1 (Nassi and Callaway (2006)).

Regarding inter cortical areas connections, MT receives inputs from many areas, such as V1, V2, V3, V3A, VP and PIP. A schema with the connections between different cortical areas is shown in Figure 3.6 (Felleman and Van Essen (1991); Born and Bradley (2005)). The 90% of the connections coming from V1 are from layer 4B (the remaining 10% connections are from layers 5 and 6), and are a majority spiny

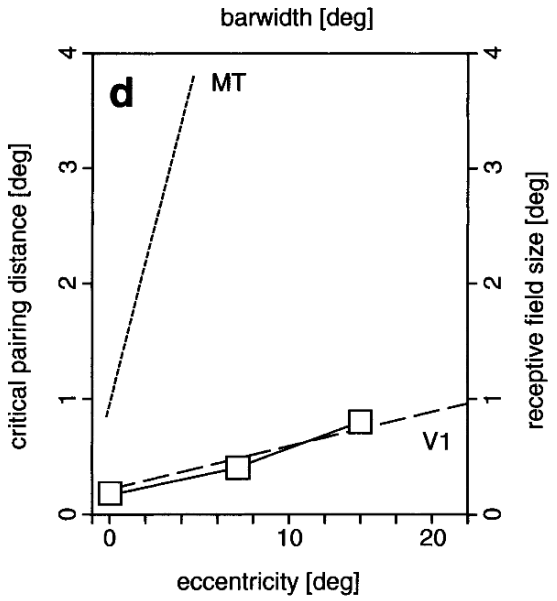


Figure 3.5: Relationship between the receptive field sizes of V1 and MT neurons versus the eccentricity. The values shown for V1 were obtained using psychophysics (Mestre et al. (2001)). The straight lines for V1 and MT were obtained from Dow et al. (1981) and Albright and Desimone (1987), respectively.

stellate cells with large cell bodies which are specialized for a fast transmission of information from the M pathway (Nassi and Callaway (2007)).

MT has also reciprocal connections with, for example the medial superior temporal area (MST), the ventral intraparietal area (VIP) and the generation of eye movements (e.g., 7a, lateral intraparietal area LIP, frontal eye field FEF, SC) (Maunsell and Van Essen (1983)). Additionally, Zaksas and Pasternak (2005) have shown that MT cells are activated by the presence of stimulus outside their receptive fields. This response is affected by the motion direction and the coherence of the motion stimuli and it has long latencies compared with conventional responses of MT cells. Their study suggests feedbacks from upper layers to higher mechanisms as e.g., attention.

3.2.2 Direction and speed selectivity

All MT cells are highly directionally-selective compared to V1 cells (Churchland et al. (2005)). Both V1 and MT layers have direction tuned neurons, but MT shows a strong inhibition in the anti-preferred direction. The proportion of directionally-selective responses is 30% in V1 and 92% in MT (Albright (1984); Snowden et al. (1991); Lagae et al. (1993)). Regarding how the MT direction-selectivity property can be created, Pack et al. (2006) found similarities between the receptive field maps of V1 complex cells and MT suggesting that MT receptive fields are primarily built by summing the outputs of V1 complex cells sharing a common preferred direction. It

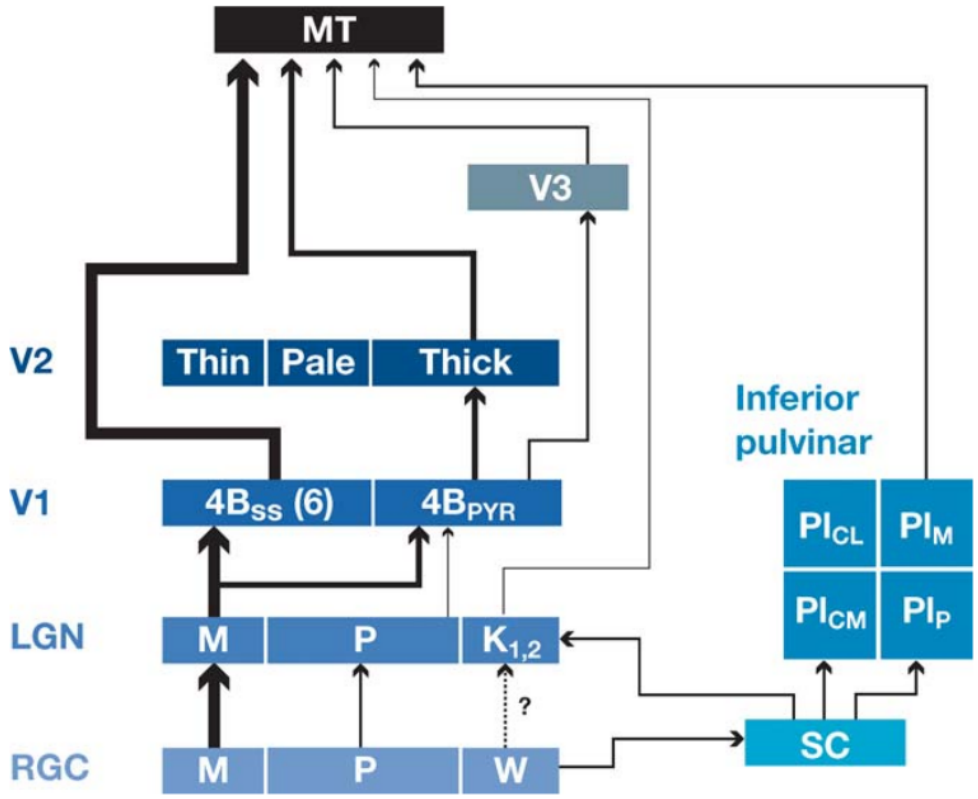


Figure 3.6: Diagram summarizing the MT input connectivity. The magnitude of inputs are directly related with line thickness (diagram taken from Felleman and Van Essen (1991)).

has been also shown that the preferred direction of a MT cell depends on the input stimulus and it evolves in time (see Section 3.2.4).

Regarding speed selectivity, Mikami et al. (1986) and Churchland et al. (2005) coincided that MT neurons tend to have higher preferred speeds compared to V1. Churchland et al. (2005) reported, for awake monkeys, a mean of 27°/s for MT neurons, while V1 neurons only have a mean of 11°/s. In anesthetized and paralyzed macaque monkeys, Priebe et al. (2006) reported a mean speed of 7.52°/s for MT neurons and a mean speed of 4.47°/s for V1 neurons, and the authors also showed an overlapping in the range of V1 and MT preferred speeds (V1: 0.3→43°/s, MT:0.4→80°/s).

Some MT cells are also tuned to speed (Maunsell and Essen (1983); Lagae et al. (1993)) [1]. The *speed-tuned* neurons are motion-sensitive cells invariant to the spatial frequency of the input stimulus and their spectral receptive fields are oriented relative to the temporal and spatial frequency axes (see Figure 3.7). Perrone and Thiele (2001) showed that a large proportion of the MT neurons tested had oriented inseparable spectral receptive fields (*speed-tuned* neurons). On the contrary, Priebe et al. (2003) found that only a small proportion of the tested cells are *velocity-tuned* neurons (27% of cells). In order to explain this difference with the previous study of Perrone and Thiele (2001), the authors suggest that this difference is due to the criteria used to assign neurons to the speed-tuned class.

Remark: *In the human brain, the spatial frequency plays an important role in speed perception. Psychophysics experiments have demonstrated that the perception of speed is influenced by the spatial frequency of the input stimulus. Low spatial frequencies bias human perception to faster speeds (e.g., Smith and Edgar (1990)).* ■

Priebe et al. (2003) also showed that the *speed-tuning* property of a MT neuron can also change with the contrast and the type of the input stimulus. Low contrast biases neurons from *speed-tuning* toward spatiotemporal independence, without altering the preferred spatial and temporal frequencies of each neuron. The type of stimulus also biased the MT responses, in the case of, e.g., square-wave gratings instead of sine-wave gratings, the responses of MT neurons were more *speed-tuned*.

The proportion of MT and V1 cells tuned for a certain speed highly differs. Lagae et al. (1993) compared the speed profiles of MT and V1 neurons for small eccentricities. They found that about 80% of V1 cells are low-pass cells while in MT is only the 55%, where ∼20% of MT cells were tuned for a higher speed. In the case of MT, the tuning properties of neurons change with the eccentricity. In small eccentricities tuned MT cells are tuned for middle range of speed (2-64 deg/s), in bigger eccentricities the range of optimal speeds is narrowed, which is consistent with the human perception measured by Orban et al. (1985).

[1] Some *speed-tuned* neurons can be also found in V1 complex cells, see Section 3.1.1

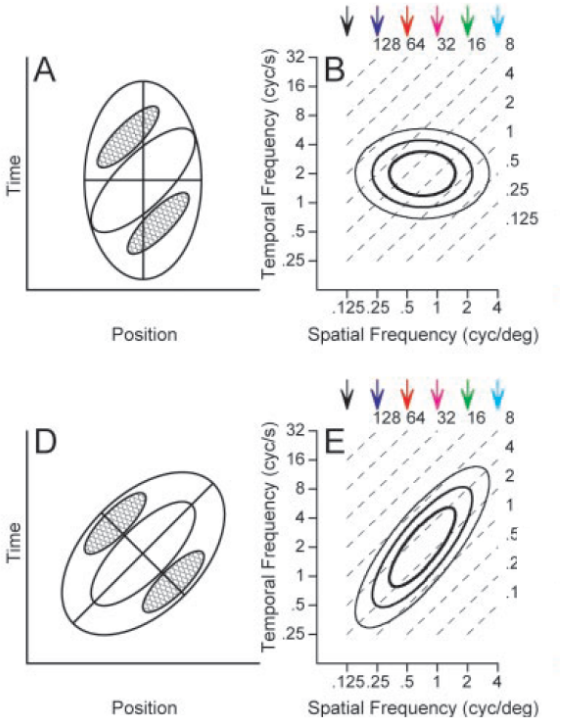


Figure 3.7: Two different models representing motion-sensitive neural responses (from Priebe et al. (2003)). A and B are motion filters where the speed tuning depends on the spatial frequency of the drifting gratings used as input stimulus. D and E are motion filter with a speed tuning independent of the spatial frequency.

3.2.3 MT surround interactions

The activation of the classical receptive field (CRF) of a MT cell is modulated by its surround. Those surrounds can be classified according to their geometry (symmetric, asymmetric) or according to the type of modulation (antagonistic or integrative). In this section we present the main characteristics.

See later: *In Chapters 7 and 8 we will show that taking into account this diversity into computational models, can provide substantial improvements in our application.* ■

Surround geometries

Different kinds of surround geometry of MT receptive fields are observed in the computation of structure of motion (see Figure 3.8). Half of MT neurons have asymmetric receptive fields introducing anisotropies in the processing of the spatial information (Lui et al. (2007)). The second half of the population examined by Xiao et al. (1997b) has two different symmetries: circular symmetry surround (20% of the population) and bilaterally symmetric surrounds, which correspond to a pair of surrounding regions on opposite sides. The neurons with asymmetric receptive fields seem to be involved in the encoding of important surfaces features, such as slant and tilt or curvature (Buracas and Albright (1996)).

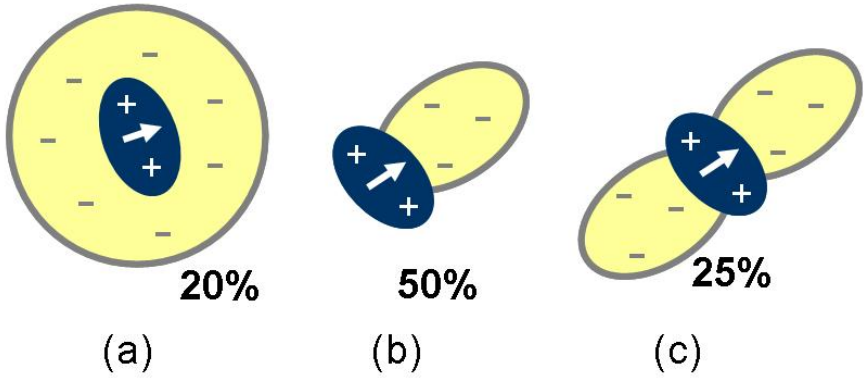


Figure 3.8: Geometries of asymmetric center-surround organization in MT cells (Xiao et al. (1997b,a)) (a) Circularly symmetric surrounds. (b) Asymmetric configuration concentrating the suppression at one side of the motion preferred axis. (c) Bilaterally symmetric zones of suppression lying in the motion preferred axis.

Surround types of modulation

Regardless the geometry of the MT receptive fields, they can be also classified according to the type of center-surround interaction as: *integrative* (for a facilitatory interaction) and *antagonistic* (for a suppressive interaction). The direction tuning

of the surround is broader than that of the center, and the preferred direction, with respect to that of the center, tended to be either in the same or opposite direction and rarely in orthogonal directions (Born (2000)). A diagram with the types of center surround interactions and their respective percentage in the neuron population studied by Born (2000) is shown in Figure 3.9. The antagonistic surrounds are insensitive to wide-field motion but very sensitive to local motion contrast. Otherwise, the integrative surrounds have better response to wide-field motion.

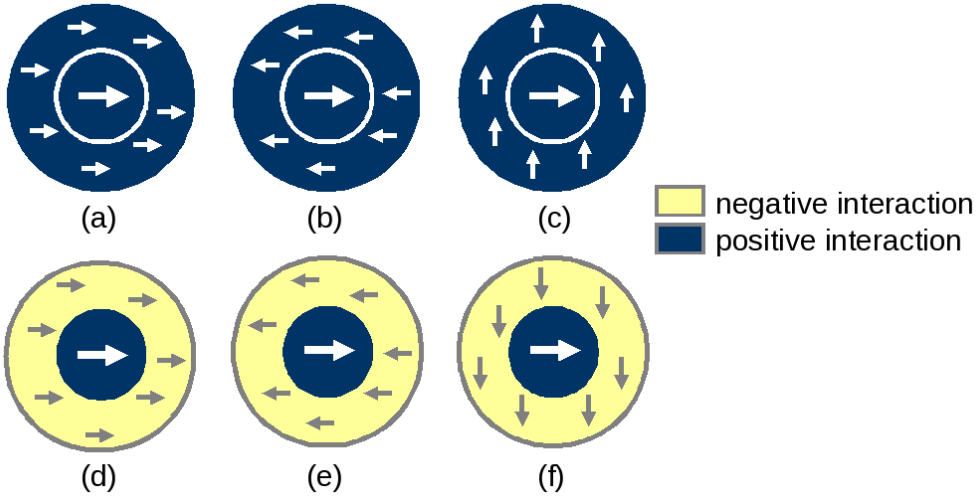


Figure 3.9: Typical interactions between the classical receptive field and its surround found by Born (2000). The surround can be either integrative ((a), (b) and (c)) or suppressive ((d), (e) and (f)). The preferred direction of the surround compared to the direction of the classical receptive field can be the same (a)-(d), opposite (b)-(e) and rarely orthogonal (c)-(f).

The interactions between center and surround can vary depending on the contrast and the type of visual stimulus (random dots, plaids, bars, etc.). Low contrasts induce an integrative surround, while high contrasts an antagonistic surround (Born (2000)). Trying different types of visual stimulus, Huang et al. (2007, 2008) showed that the surround can be integrative if the motion information contained inside the classical receptive field of the MT cell is ambiguous (aperture problem). On the contrary, the surround acts as a segmentation (antagonistic surround) if the motion direction perceived inside the classical receptive field is solved.

3.2.4 Preferred direction of MT cells

The preferred direction (PD) of a MT cell has been generally measured through a drifting grating, where most of the times the cell shows a clear direction selectivity. However, the behavior of a MT neuron highly depends on the input stimulus, the contextual information and time. This dependency changes the type of surround modulation, as it was mentioned in Section 3.2.3, and therefore, the PD of a MT cell.

Time dependency of MT preferred direction

Several studies, such as Pack and Born (2001); Pack et al. (2004) and Born et al. (2006) showed that the PD can be modified depending on the input stimulus. Specifically, Pack et al. (2004) showed that the PD measured using barberpoles instead of grating is biased toward perception, i.e., the side of the barberpole with the longest side. This PD deviation, compared to the one measured drifting grating, depends on the aspect ratio of the barberpole (see Figure 3.10).

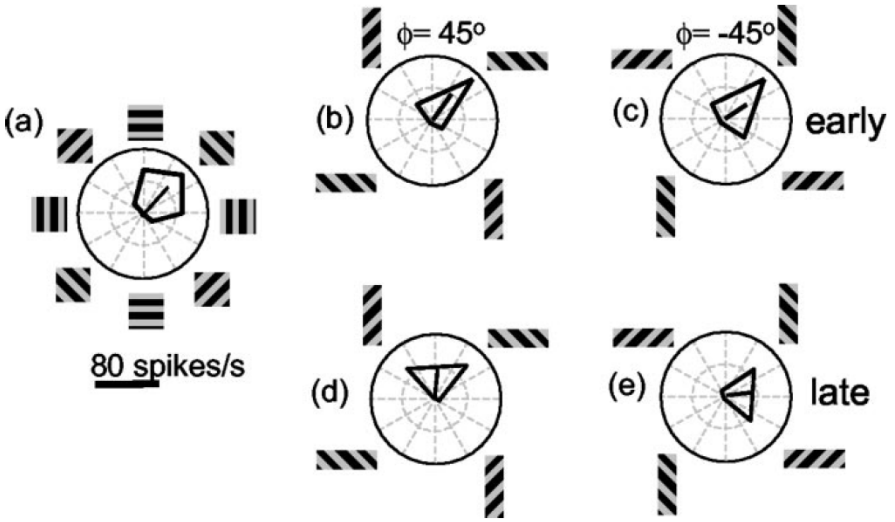


Figure 3.10: Response of a MT cell to a drifting grating (a) and to a barberpole (b)-(c)-(d)-(e) with different orientations (image taken from Pack et al. (2004)). The preferred direction (PD) of a MT cell measured with gratings is shown in (a). (b)-(c)-(d)-(e) show how the PD of the MT cell moves from a similar value than the one measured with drifting gratings (*early*) towards a value related with perception (*late*).

Evidence of microelectronic recordings in MT of alert monkey reveal that during the first 80ms after the onset stimulus the response is strongly biased by 1D motion, i.e., the direction defined by the orthogonal direction to the contours, but lately the 2D motion direction is encoded. These experiments suggest that the aperture problem is solved within the first 100ms of the onset stimulus (Pack and Born (2001)).

The mechanisms underlying the PD deviation of a MT cell are unknown. It looks like that the primate visual system initially considers all the information available (ambiguous and unambiguous), and that along time, it refines it in order to solve the aperture problem. This convergence in time can be associated to different and complex neural networks which convey information coming from other areas of the visual system as feedbacks (Berzhanskaya et al. (2007)) or horizontal connections. This phenomenon is also associated to the contribution of terminators or end-points in different areas of the visual field such as V2 or V1 (Berzhanskaya et al. (2007); Bayerl and Neumann (2007); Pack et al. (2003)) which should require slightly longer latencies.

See later: *In Chapter 9 we will study the effect of V1 surround inhibition in the PD of MT cells.* ■

Another example of how the response of a MT cell changes with the input stimulus is the work done by Huang et al. (2007, 2008) who showed that surround modulation in area MT can be either antagonistic or integrative depending on the visual stimulus context, changing by this way the perceptual interpretation of the input stimuli. Most of the previous experiments performed on MT surround interactions found that surround inhibits the activation of the classical receptive field (CRF) (antagonistic surround, e.g., Xiao et al. (1997b)). But, Huang et al. found that only the motion information coming from the same object induces integration. Motion signals coming from different objects should be segregated to achieve segmentation. The surround integration previously reported by Born (2000) in owl monkeys is not the same case. The directional reinforcement is not surround modulation, because it induced responses even in the absence of the stimuli in the CRF.

Pattern and component cells

Comparing the preferred direction of MT neurons for gratings and plaids, it is possible to classify them as *pattern* direction selective (PDS) or *component* direction selective (CDS). The PDS neurons have a unimodal response for plaids, while the CDS neurons show a bimodal response indicating the two directions of the gratings conforming the plaid stimulus (see Figure 3.11). The response of a PDS cell to a plaid is generally quite different from the sum of its responses to the individual gratings alone.

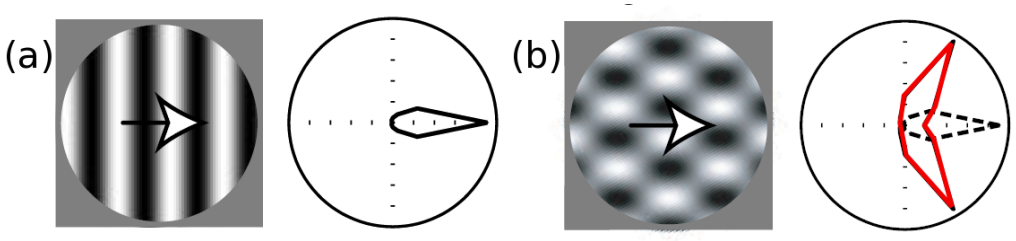


Figure 3.11: Figure shows typical responses of MT cells to drifting gratings (a) and plaids (b). The MT cell response for a plaid stimulus (b) could be either sensitive to the true direction of motion (*PDS* cell, black dashed lines) or sensitive of the components conforming the plaid (*CDS* cell, red lines).

The fact that the time response of CDS neurons is faster (about 6ms) than PDS neurons (about 50-75ms), suggests a two-stage model for MT, where the outputs of the CDS neurons are used as inputs of the PDS (Movshon et al. (1986)). The selectivity of a PDS cells evolves during the first 100-150 ms after the exposition of a complex stimulus as plaid (Smith et al. (2005)), starting with a broader selectivity resembling

CDS cells. After some tens of milliseconds, their responses evolve to be more PDS-like.

On the other hand, CDS cells give a stable response as soon as the stimulus is set. A diagram of the evolution of cells along time can be found in Figure 3.12. The proportion of cells belonging to PDS or CDS found by Smith et al. (2005), for anesthetized monkeys, is ∼41% for CDS, ∼25% for PDS and ∼34% unclassified. These values are similar with the ones found in awake (Stoner and Albright (1992)) and anesthetized (Rodman and Albright (1989); Priebe et al. (2003)) monkeys using the same stimuli.

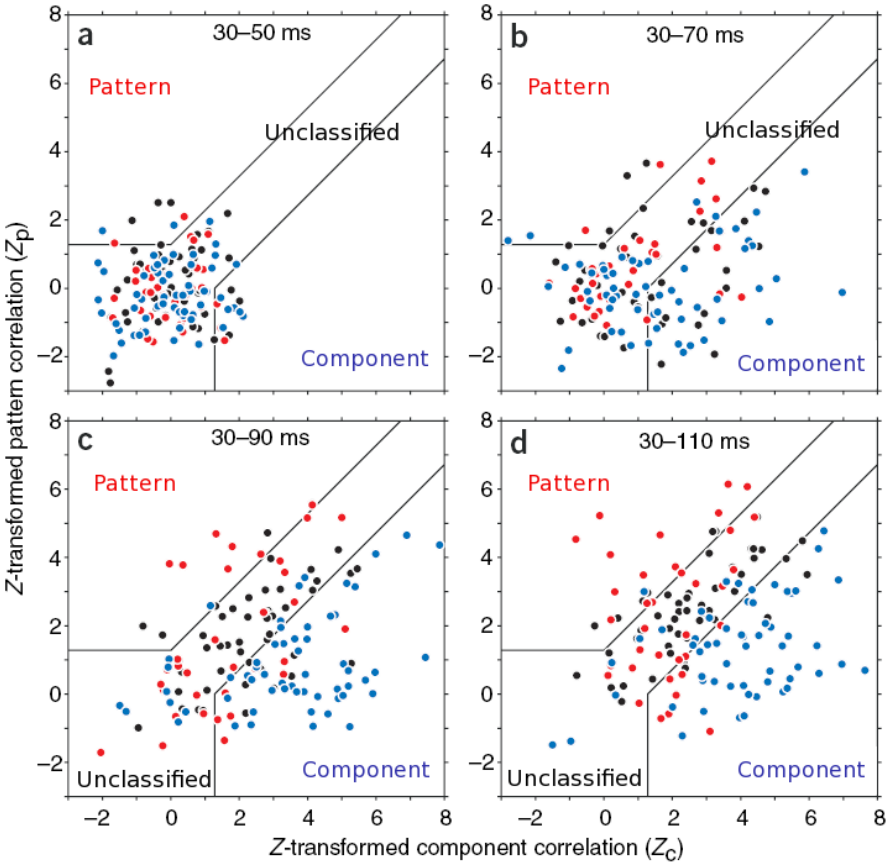


Figure 3.12: Red dots represent PDS cells, blue dots represent CDS cells and black dots represent cells which are not classified. Three areas are represented in the diagram, labeled as *Pattern*, *Component* and *Unclassified*. The figures shows the evolution of the cell classification over time with a cumulative window starting at 30-50 ms and finishing at 30-110 ms. It is possible see than CDS cells are classified sooner than PDS cells (from Smith et al. (2005)).

Is the pattern-motion computation (PDS cells) done locally or globally inside the MT receptive field? Majaj et al. (2007) asked whether MT cells just pooled information from V1 neurons treating this input information as a single object. They found that an important computation of the PDS mechanism is done locally inside small patches of the MT receptive field (in a finer scale than the whole MT receptive field), suggesting that spatial coincidence of the components of a moving object is required. In Perrone and Krauzlis (2008), a MT model is proposed to explain this phenomenon.

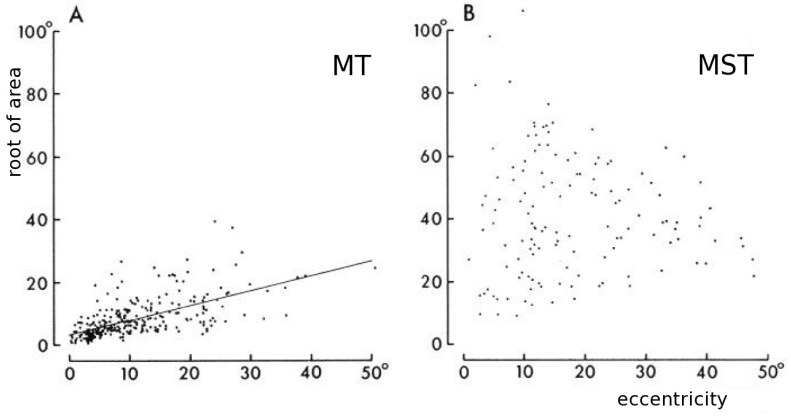


Figure 3.13: Receptive field size of MT neurons and MSTd neurons sensitive to planar motion. The square root of the excitatory area of the receptive field is plotted versus the eccentricity (image taken from Tanaka et al. (1986)) .

3.3 MST: THE MEDIAL SUPERIOR TEMPORAL AREA

MST is above to MT in the visual motion hierarchy. MST receives connections mainly from MT (Maunsell and Van Essen (1983)) and, equally to MT, it is located at the superior temporal sulcus (STS).

MST has at least two main divisions:

- MSTd: dorsal part localized in the anterior bank of the STS with large receptive fields. Neurons in this region are directionally-selective for moving visual stimuli responding to flow-field stimuli as (Saito et al. (1986)): rotation, radial (expansion/contraction) and planar motion (see Figure 3.14). Cells in this area have large receptive fields irrespective of the eccentricity (see Figure 3.13) with a mean value of 41° (Tanaka et al. (1986)).

- MSTl: ventral-lateral region of the medial superior temporal area with small receptive fields similar to MT size. These neurons are important for the self-motion compensation and smooth pursuit eye movements (Churchland and Lisberger (2005); Inaba et al. (2007); Ilg (2008)).

In MSTd, originally Saito et al. (1986) and Tanaka and Saito (1989) found three classes of directionally selective cells, each of them responding for a certain flow-motion pattern: planar motion (51%), radial motion (16%) and rotation (14%). Lately, Duffy and Wurtz (1991) showed that neurons in MSTd are able to combine the different motion patterns finding that neurons responding only to one flow-motion pattern were only the 23% of the population studied. The remaining 34% and 29% responded to two components (plano-circular or plano-radial but never circulo-radial) and three components, respectively.

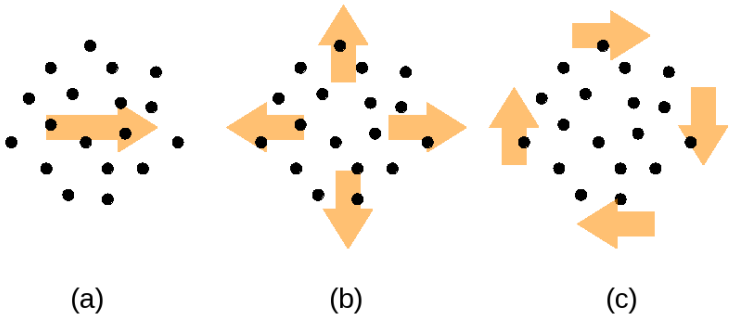


Figure 3.14: Different flow-field patterns detected by MSTd cells. (a) planar motion, (b) radial motion (expansion represented) and (c) rotation motion (clockwise direction represented).

In order to better characterize MSTd neurons, several studies combining different motion aspects have been done. Graziano et al. (1994) found that many MSTd neurons are preferentially selective to spiral motion. They also found that MSTd cells selectivity is maintained inside its large receptive field. By the other hand, Geesaman and Andersen (1996) asked whether the direction tuning of MSTd neurons is form/cue invariant inside its receptive fields. The experiments were performed using coherently moving random dots, solid squares, outlines of squares and squares of stationary random dots. These experiments did not revealed a significant variance in the direction tuning of MSTd neurons. Studies done by Duffy and Wurtz (1997) also showed that changes in the gradient of speed of the flow-field patterns (slower speed at the center and faster speed in the periphery) highly alters the response of MSTd neurons.

Regarding latencies, Kawano et al. (1994) used large field random dot patterns to measure the response latencies of MST neurons. They found that 80% of neurons were activated during the first 50ms after the stimulus onset. They also measured the latency of ocular responses finding that both neuronal and ocular responses decreased as stimulus speed increased. The time differences between neuronal response and ocular response varied little with stimulus speed.

CHAPTER 4

MOTION MODELS

"I can calculate the motion of heavenly bodies, but not the madness of people."
–Isaac Newton (1643-1727)

Contents

OVERVIEW

Along years many models of motion processing have been proposed in several communities such as computer vision or neuroscience. Most of the bio-inspired models proposed attempted to understand the role and interactions between the different brain areas involved in visual motion processing. Within the bio-inspired methods, just a few have been proposed in order to be also applied in real applications.

The first part of this chapter will introduce the main motion detection techniques which are the first step of any motion model. The motion detection techniques can be divided into three main categories: differential methods, frequency-based methods and region-based matching methods. Considering our contribution, we will focus here on the first two categories..

The second part of this chapter summarizes the bio-inspired motion models present in the literature, such as e.g., Bayerl and Neumann (2004), Grossberg et al. (2001), Perrone (2004), Simoncelli and Heeger (1998) or Nowlan and Sejnowski (1994)-Nowlan and Sejnowski (1995).

Keywords: motion detection, optical flow, spatiotemporal filtering, motion models, V1, MT.

Organization of this chapter:

This chapter is organized as follows: Section 4.1 describes the different family of approaches proposed for motion detection. Within them, two are described in detail: differential techniques (Section 4.1.2), and frequency-based methods (Section 4.1.3). Section 4.2 reviews the motion models inspiring the development of this thesis, grouping them into: feedforward models (Section 4.2.2) and recurrent models (Section 4.2.3).

4.1 MOTION DETECTION

4.1.1 Three main categories

Motion detection in video sequences has been widely studied since the last 20 years by different communities, such as, computer vision, robotics, biological vision, signal processing, etc. All the techniques present in the literature can be divided into three maincategories (Simoncelli (1993); Barron et al. (1994)):

1. **Differential techniques:** Also known as "gradient" techniques, estimate the optical flow vectors from the derivatives of image intensity over space and time.

2. **Frequency-based methods:** The most important category in the development of this thesis. These methods are based on spatiotemporal oriented filters and motion is treated in the frequency space. These frequency-based filters are also divided into two other categories: *energy-based* filters and *phase-based* filters.

3. **Region-based matching**: These techniques attempt to match "features" (small regions of the image) from frame to frame. The matching criterion is usually least squares or normalized correlation measure.

In the rest of this chapter, we will comment further the first two categories (see Borst (2007) for a comparative study). The latter one, namely the region-based techniques, will not be considered because its conception is far from biological plausibility.

4.1.2 Differential techniques

Differential techniques operate over the assumption that the intensity of the image is preserved over time (*brightness change constraint*). Changes in image intensity are only due to translation of the local image intensity and not to changes in lightning, contrast, etc. Under this assumption, the derivative of the image intensity $L(\mathbf{x}, t)$ with respect to time t must be zero for each point $\mathbf{x}(t) = (x(t), y(t))$ along its trajectory t, i.e.

$$\frac{dL(\mathbf{x}(t), t)}{dt} = 0. \tag{4.1}$$

Deriving equation (4.1) gives

$$\frac{\partial L}{\partial x} v_x + \frac{\partial L}{\partial y} v_y + \frac{\partial L}{\partial t} = 0, \tag{4.2}$$

where $\mathbf{v} = (v_x, v_y)^T = \left(\frac{dx}{dt}, \frac{dy}{dt}\right)$ is the instantaneous optical flow vector (see Fleet and Weiss (2005) for further details).

Equation (4.2) is called the *optical flow contraint*. The solutions of (4.2), plotted in the velocity space, define the constraint line. The constraint line represents all the

2D velocities that are consistent with image derivative. Equation (4.2) also defines a conservation law which is only true for rigid translation of a Labertian surface in the image plane. Equation (4.2) can be written in the following more compact form:

$$\nabla L \cdot \mathbf{v} + L_t = 0, \tag{4.3}$$

where ∇ is the gradient symbol and $L_t = \frac{dL}{dt}$.

Equation (4.2) is a single linear equation with two unknowns that cannot be fully solved because it has not an unique solution. This is an *ill-posed* problem known as the *aperture problem*, where equation (4.2) gives as solution a family of velocities along a line in the velocity space. This line is perpendicular to ∇L, and its perpendicular distance to the origin is given by $|L_t|/\|\nabla L\|$. Equation (4.3) must be constrained in order to find an unique solution.

Several solutions were proposed to solve equation (4.3). Let us mention some of them:

- Apply second-order differential methods. The second-order differential methods, derived from the conservation of $\nabla L(\mathbf{x}, t)$, $d\nabla L(\mathbf{x}, t)/dt = 0$, use second-order derivatives to contrain 2D velocity

$$\begin{bmatrix} L_{xx} & L_{yx} \\ L_{xy} & L_{yy} \end{bmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} + \begin{pmatrix} L_{tx} \\ L_{ty} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.4}$$

  These methods have stronger restrictions than the ones needed for Equation (4.3) and several possibilities have been proposed for its solution (see Otte and Nagel (1994); Tistarelli (1995)). In this case the first-order deformations of intensity, such as rotation or dilatation, should not be present. How the 2nd order derivatives are sometimes hard to measure, the 2nd order methods are normally less accurate than estimates from 1st order methods.

- Write gradient constraints from nearby pixels, assuming that they share the same 2D velocity. The solution will be the velocity v which minimizes the contraint errors. To do this, the least-square (LS) estimator is used:

$$E(\mathbf{v}) = \sum_{\mathbf{x}} g(\mathbf{x}) \left[\nabla L \cdot \mathbf{v} + L_t\right]^2, \tag{4.5}$$

  where $g(\mathbf{x})$ is a weighting function which defines the region where the constraints will be applied, normally a Gaussian. This approach, which gives good results, has been extended and it is often used in computer vision applications. However, it is local, and there is no notion of global regularity for the resulting flow.

- One may also use parametric models of velocity that respect as much as possible the optical flow constraint. In the affine case, one looks for σ such that

$$\mathbf{v}(\mathbf{x}) = \mathbf{v}_\theta(\mathbf{x}) = \begin{pmatrix} \theta_1 + \theta_2 x_1 + \theta_3 x_2 \\ \theta_4 + \theta_5 x_1 + \theta_6 x_2 \end{pmatrix},$$

where the unknown parameter vector $\theta \in \mathbb{R}^6$ is determined by minimizing

$$E(\theta) = \int_{\Omega} \phi(\nabla L \cdot \mathbf{v}_\theta + L_t) d\mathbf{x},$$

where ϕ is a suitable given function. Within the models proposed to solve this optimization problem one can mention Irani and Peleg (1993) and Odobez and Bouthemy (1995).

- Regularizing the velocity field is another possibility, where the idea is to minimize

$$\inf_{\sigma} \left(A(\sigma) + S(\sigma) \right), \tag{4.6}$$

where $A(\sigma)$ is the fidelity term and $S(\sigma)$ is the smoothing term.

Horn and Schunck (1981) (see also Schnörr (1991)) were the first to solve this regularization optimization problem. Then, many solutions were proposed: modifying the Horn and Schuck functional (Black (1992); Black and Rangarajan (1996); Nési (1993)), adding some penalties based on divergence and the rotational of the flow field (Suter (1994); Gupta and Prince (1996); Guichard and Rudin (1996)), proposing an oriented smoothness constraint in order to handle occlusions (Nagel (1983, 1987, 1989); Enkelmann (1988)), among others.

4.1.3 Frequency-based methods

Frequency-based methods treat motion in the Fourier domain, where the optical flow is obtained filtering the input image in space and time (*spatiotemporal* filtering). Methods can be classified in:

- **Phase-based:** The velocity is defined in terms of the phase behaviour of the outputs of the spatiotemporal filters (see, e.g., Fleet and Jepson (1990)).
- **Energy-based:** These techniques use the energy of the output of the spatiotemporal filters to compute the optical flow (see, e.g., Adelson and Bergen (1985)).

The spatiotemporal filtering techniques are physiologically motivated and most of them come from the computational biology community. In the spatiotemporal $(\mathbf{x}, t)$ space, the problem of detecting motion becomes a problem of detecting spatiotemporal orientation. The spatiotemporal oriented filters interpret motion as an orientation in $(\mathbf{x}, t)$ space. For example, Figure 4.1(a) shows two vertical bars moving continuously from left to right at different speeds. Considering only undimensional motion (x-axis only), the two-dimensional spatiotemporal diagram is represented in Figure 4.1(b), where the motion becomes a slanted bar. The slant reflects the velocity of the motion ($v_g > v_b$).

The spatiotemporal filters share many properties with V1 motion detectors and they have been extensively used in the litterature

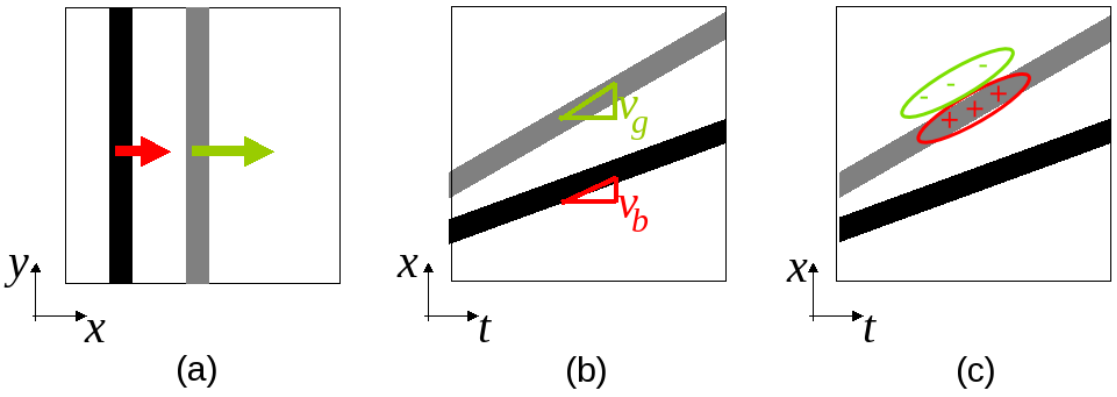


Figure 4.1: A black and a gray bars are moving from left to right in a uniform background (a) with different speeds v_b and v_g, respectively, where v_g is faster than v_b. The orientation in the space-time (x, t) for both bars is shown in (b) where the slant fits the speeds v_b and v_g. Higher speeds bring out higher slants. How the motion is unidirectional, the representatio in the space-time (y, t) is ommited. (c) Spatiotemporal oriented filter that can be used to detect the velocity of the gray bar.

(Hubel and Wiesel (1962); Watson and Ahumada (1983); Adelson and Bergen (1985); Watson and Ahumada (1985); Van Santen and Sperling (1985); Fleet and Jepson (1989); Simoncelli and Heeger (1998); Grzywacz and Yuille (1990); Ringach (2002); Conway and Livingstone (2003)).

In general, they are modeled as Gabor filters oriented in the $(\mathbf{x}, t)$ space. To detect a motion profile in the $(\mathbf{x}, t)$ space, the orientation of the Gabor filter should coincide with the $(\mathbf{x}, t)$ orientation of the input stimulus (see Figure 4.1 (c)).

But interpreting the output of a spatiotemporal filter is not an easy task. Its response varies in time and highly depends on the contrast and luminance level of the input stimulus. For a drifting sinusoidal grating, the spatiotemporal filter output will be also a sinusoidal with an amplitude and phase related to the input grating. So, we cannot use directly the instantaneous value of those filters as motion quantifiers.

The energy filters, proposed by Watson and Ahumada (1983, 1985), Van Santen and Sperling (1985) and Adelson and Bergen (1985) tackle some of the problems present in spatiotemporal filtering. They proposed an energy filter which combines the ouputs of spatiotemporal filters with different phases. The following sections briefly describe the solutions proposed by Watson and Ahumada (1985), Van Santen and Sperling (1985) and Adelson and Bergen (1985), showing that these three proposals share the same phylosophie.

Watson and Ahumada motion detectors

Watson and Ahumada (1983, 1985) are one of the first studies of motion in the frequency domain. They analyzed the frequency spectra of moving images, proposed simple solution to long-standing problems in motion perception and proposed a linear motion sensor as a motion detector candidate.

Their motion sensor considers V1 simple cells properties of visual cells in the cortex of the cat and monkey. These simple cells can be modeled by a 2D Gabor filter where the diameter of the surrounding Gaussian at half height is 1.324 times the period of the sinusoid. This diameter gives a spatial frequency bandwidth (at half height) of one octave.

The motion sensor, as a spatiotemporal filter, will be constructed considering a spatial filter entity $g(x, y)$ and a temporal filter entity $f(t)$.

The basic spatial filter $g(x, y)$ is defined as

$$\begin{aligned} g(x,y) &= a(x)b(y) \\ a(x) &= \exp\left(-x^2/\lambda^2\right)\cos(2\pi u_s x) \\ b(y) &= \exp\left(-y^2/\lambda^2\right), \end{aligned} \tag{4.7}$$

where u_s is the frequency of the cosine and λ the spread of the Gaussians.

The basic temporal filter $f(t)$ is modeled with a biphasic profile obtained by the impulse response

$$f(t) = \xi\left[f_1(t) - \zeta f_2(t)\right], \tag{4.8}$$

where

$$f_i(t) = \frac{\Theta(t)}{\tau_i(n_i-1)!}(t/\tau_i)^{n_i-1}\exp(-t/\tau_i), \tag{4.9}$$

$\Theta(t)$ is the Heaviside function, and ξ, ζ, τ_i, n_i are constants. The shape of the temporal profile is shown in Figure 4.2.

The response of the motion sensor is then obtained convolving the impulse response of all the cascade elements and adding their responses in parallel (process summarized in Figure 4.3).

New entities are shown in the block diagram,

- The Hilbert spatial filter $h(x) = -1/\pi x$ converts odd functions into even, and even into odds. Two functions that are Hilbert transformed of each other are said to form a quadrature pair.

- The temporal delay $\delta(t - \tau)$.

- The Hilbert temporal response $h(t) = -1/\pi t$ also converts the temporal impulse response to a quadrature version of the original input.

This linear motion sensor remains as a linear filter which simulates the responses of V1 simple cells but not V1 complex cells. The next solution presented by Adelson and Bergen shows how nonlinear combination of simple cells can generate a V1 complex cell.

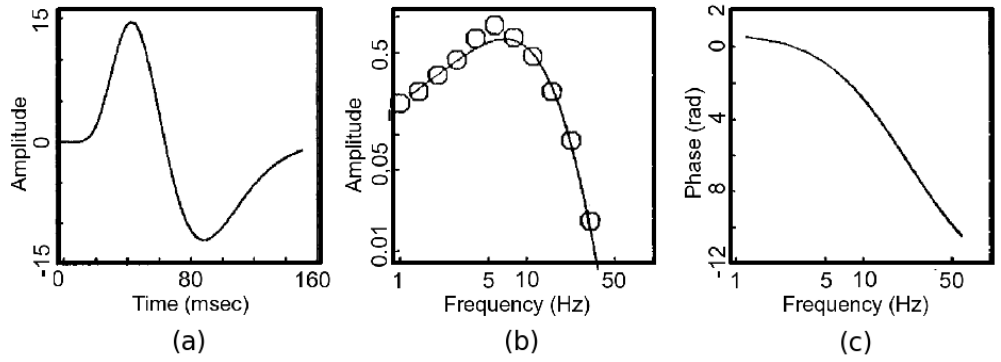


Figure 4.2: Temporal profile of the motion detector presented by Watson and Ahumada (1983, 1985): (a) impulse response, (b) amplitude response and (c) phase response (image taken from Watson and Ahumada (1983)).

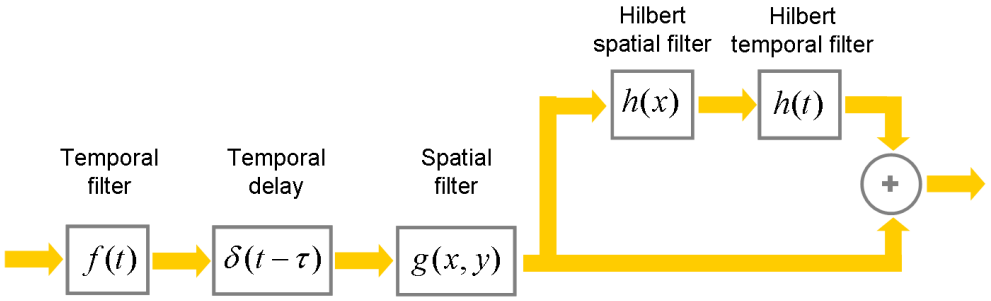


Figure 4.3: Block diagram of the linear motion sensor proposed by Watson and Ahumada (1985).

Adelson and Bergen energy filters

As we previously stated, spatiotemporally linear filtering presents two main issues. First, it is phase sensitive, e.g., its response sign depends on the constrast of the input stimulus. Second, it is not possible to use its instantaneous response as a simple measure of motion, e.g., for input drifting gratings we obtain oscillating responses. So, a more complicated process is needed in order to have a measure of motion independent of the polarity and instantaneous phase of the input stimulus.

To do so, a phase-independent motion detector, proposed by Adelson and Bergen (1985) (see Figure 4.4). Two linear spatiotemporal units are combined summing their squared responses. The linear spatiotemporal units are in quadrature, i.e., with a different of phase of 90°. For mathematical convenience the ideal case of Gabor functions is considered, one with an even phase (cosine) and the other with an odd phase

(sine). Typical unidimensional Gabor spatiotemporal units are:

$$F^{odd}(x,t) = \exp\left(-\frac{x^2}{2\sigma^2}\right)\sin\left(\omega_x x + \omega_t t\right) \tag{4.10}$$

$$F^{even}(x,t) = \exp\left(-\frac{x^2}{2\sigma^2}\right)\cos\left(\omega_x x + \omega_t t\right), \tag{4.11}$$

where σ is the standard deviation of the surrounding Gaussian, ω_x and ω_t are the spatial and temporal frequencies, respectively.

The local motion energy is extracted squaring and summing the two units outputs. Since the two Gabor functions are sine and cosine functions weighted by a Gaussian window, the energy is extracted inside a spatiotemporal frequency band.

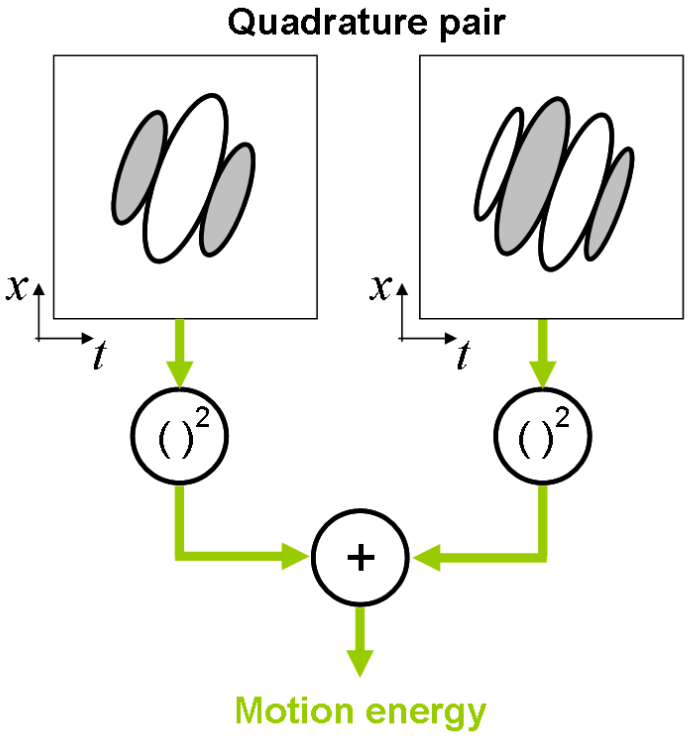


Figure 4.4: Adelson and Bergen (1985) energy filters: Two linear filters in quadrature, i.e., with responses 90°out of phase are combined to create an energy motion detector. The energy motion detector created is phase-independent (for a given spatial-frequency band) and is obtained summing the squared response of each linear filter.

The energy filter obtained is phase-independent and for a constant rightward motion of a drifting grating will give an unmodulated positive response. Its response is independent to the sign of the contrast, but it depends on the amplitude of the contrast.

A weak response can be interpreted as two different events: an object moving too slow or too fast, or an object moving with an intermediate speed but with a low contrast. So, contrast and speed are mixed up. In a following work, Adelson and Bergen

(1986) proposed a mechanism to extract a signal related to speed independently of constrast:

$$v = \frac{\left(R^2_{odd} + R^2_{even}\right) - \left(L^2_{odd} + L^2_{even}\right)}{\left(S^2_{odd} + S^2_{even}\right)}, \tag{4.12}$$

where R, L and S denote the output of filters tuned for righward motion, leftward motion and stationary stimuli, respectively.

Elaborated Reichardt motion detectors

The Reichardt detector, originally proposed by Reichardt (1957), attempted to model fly visual system. Reichardt model has been the source of inspiration of many approaches such as the ones proposed by Van Santen and Sperling (1984, 1985) and Bayerl and Neumann (2004).

Reichardt assumed that motion detectors are composed of two subunits tuned to motion in opposite directions (left and right). The subunit tuned to leftward motion is substracted to the subunit tuned to rightward motion, and vicecersa. For example, if the output of the left unit exceeds the output of the right unit, then the motion detector will indicate the leftward direction. Analogously, if the output of the right subunit exceed the output of the left subunit, the motion detector indicates rightward motion.

But, the motion detector proposed by Reichardt (1957) presented spatial and temporal inconvenient aliasing, specially for certain choices of the temporal filter which could lead to incorrect direction responses. To overcome this difficulty, Van Santen and Sperling (1985) proposed an Elaborated Reichard Detector (ERD) which correctly indicates motion direction for any spatial and temporal frequency (Figure 4.5).

The aliasing problem is eliminated chosing the right spatial and temporal filters (SF and TF) to have the correct sign of the detector output. This detector, known as Elaborated Reichardt Detector (ERD), also assumes that only the final response is used, i.e., the difference between the two subunit responses. Using specific assumptions for the temporal filter TF and the spatial filter SF, the response of the subunits can be equivalent or even better than the response of the whole detector. For this, the authors proposed two variants:

1. Subunits where the temporal filter delays all the temporal frequencies of the subunit by a quarter of a temporal cycle ($\pi/2$ temporal phase-delay subunits).

2. Subunits where their input receptive fields have the properties: for every spatial frequency, the spatial positions of a sinusoidal grating that maximize the response of the receptive fields differ by one quarter of a spatial cycle ($\pi/2$ spatial phase shift subunits).

The motion detector proposed by Watson and Ahumada (1983) shares the same structure, but not components, than ERD. Basically, Watson and Ahumada (1983)

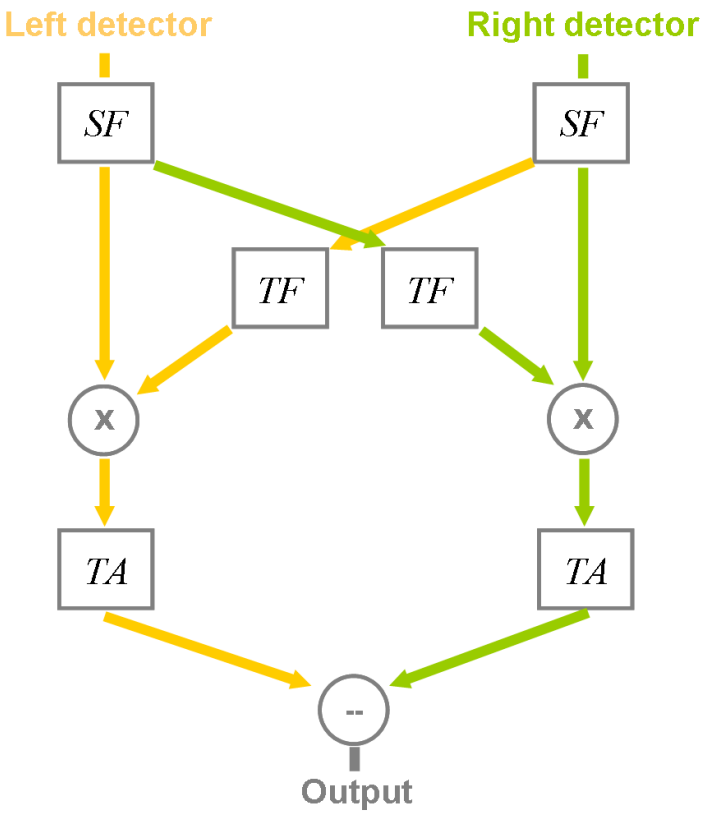


Figure 4.5: The Elaborated Reichardt Detector (ERD) proposed by Van Santen and Sperling (1985). SF are linear spatial filters with spatial response r_{right} and r_{left}, respectively; TF indicates a linear time-invariant filter; $\times$ indicates a multiplication unit; TA indicates a temporal integration operation, and $-$ indicates a unit that substracts its left (orange path) from its right (green path) input.

replace the ERD' s multiplier by an adder assuming a temporal delay of $\pi/2$ and a spatial phase-sift of $\pi/2$. Van Santen and Sperling (1985) showed that their ERD can be transformed to act as a Watson and Ahumada (1983) motion detector.

Van Santen and Sperling (1985) also showed that the ERD is equivalent to Adelson and Bergen (1985) motion detector, both motion detectors perform the same operations in a different sequence. At subunits level, in the case of $\pi/2$ property for the temporal delay and spatial receptive field filters, Adelson and Bergen (1985) outputs differ from ERD outputs by an additive constant K.

Reichardt motion detectors are ideal under low luminance conditions where noise is prominent. Whereas the gradient detectors show a superior performance under high luminance conditions where signal-to-noise levels are high (Potters and Bialek (1994); Borst (2007)).

WIM sensor

The *WIM* (Weighted Intersection mechanism Model) sensor, proposed by Perrone and Thiele (2001); Perrone (2004), sensitive to a certain speed is built up in stages from two spatiotemporal filters with properties based on V1 neurons.

In primates, some V1 neurons act as low-pass filters while others as band-pass filters. Low-pass neurons better respond to static patterns, while band-pass neurons prefer moving features. The terms *sustained* and *transient* have been assigned low-pass and band-pass neurons, respectively. *Sustained* indicates a response that extends for the duration of the stimulus (low-pass neurons), while *transient* indicates a response primarly at stimulus onset and offset (band-pass neurons).

By combining the outputs of two spatiotemporal filters (one non-directional (sustained) and another directional type (transient)) the authors defined a *weighted intersection mechanism* WIM that produces an elongated and oriented filter in the spatiotemporal frequency domain. This mechanism enables two filters with broad temporal tuning (one low-pass and the other band-pass) to be converted into a filter with tight temporal frequency tuning and an orientation that maps onto the oriented spectra generated by moving edges. The authors also showed that the speed tuning property of such a WIM filter is comparable to that found in many MT neurons. However, their analysis was only restricted to the speed tuning properties of the filter and they did not discuss the direction tuning of their motion sensor.
The WIM sensor definition is based on the two following principles

1. The maximum output from the new speed tuned mechanism occurs whenever the outputs of the transient and sustained neurons are equal.

2. The peak response of the speed-tuned mechanism is maximal only for specific edge speeds, i.e., for spatial and temporal frequency combinations that lie along the oriented line in frequency space.

Examples of the velocity-tuned neurons (motion detectors) obtained by the *WIM* mechanism are shown in Figure 4.6
Following these two principles, a WIM sensor tuned to speed v, i.e., to all spatiotemporal frequencies combinations (u, ω) such that $v = \omega/u$, is defined as

1. The two V1 contrast sensitivity neurons (S: sustained and T: transient) can be defined to that they overlap along the $v = \omega/u$ line by modifying the spatial frequency tuning of the transient neuron relative to the sustained neuron.

   Temporal frequency contrast sensitivity tuning z_t, for each value of the stimulus temporal frequency ω, is modeled and fit using the following equation

   $$z_t(\omega) = \sqrt{m_1^2 + (\zeta^2 + m_2^2) - 2\zeta m_1 m_2 \cos(\vartheta_1 + \vartheta_2)}, \tag{4.13}$$

   where $m_1 = ((2\pi\omega\tau_1)^2 + 1)^{-9/2}$, $m_2 = ((2\pi\omega\tau_2)^2 + 1)^{-10/2}$, $\vartheta_1 = -9\arctan(2\pi\omega\tau_1)$ and $\vartheta_2 = -10\arctan(2\pi\omega\tau_2)$. τ_1 and τ_2 are time constants measured in seconds.

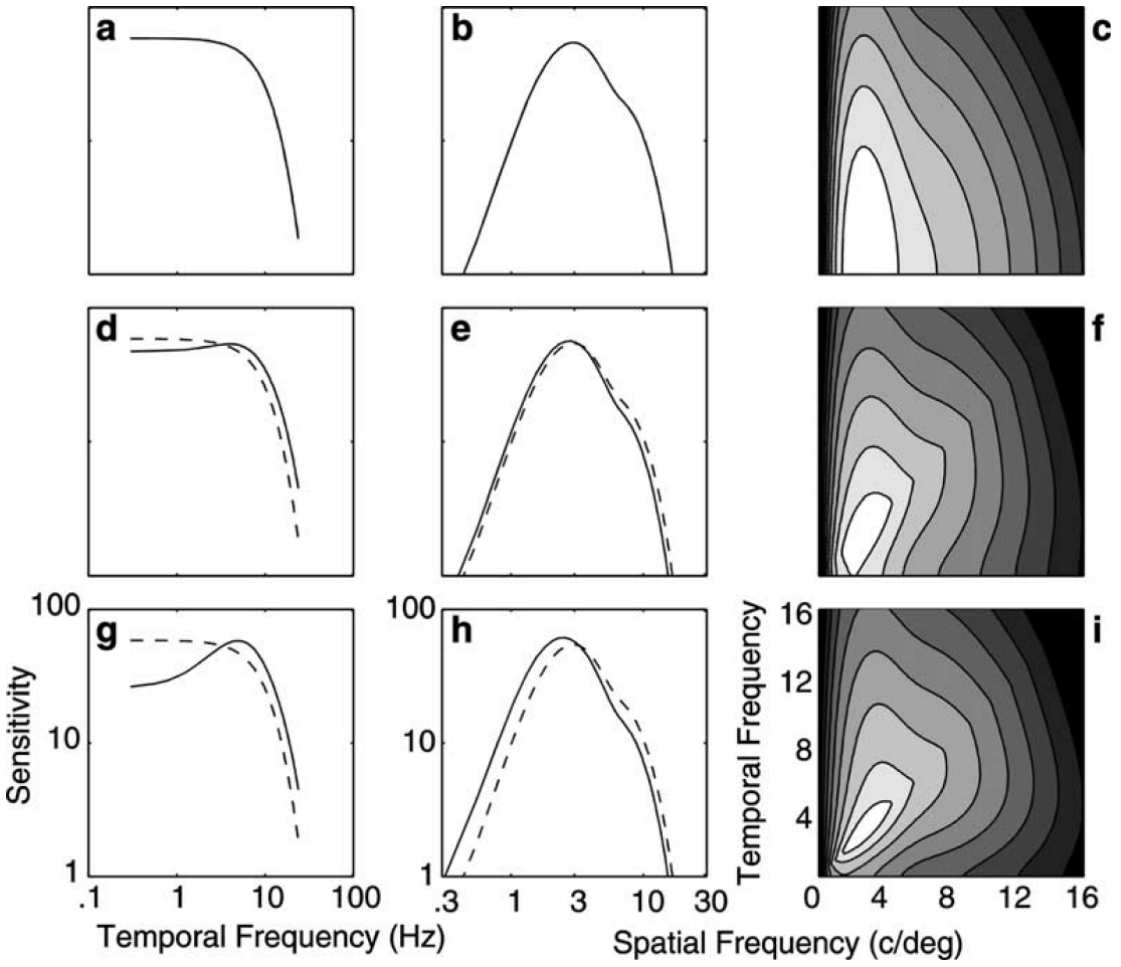


Figure 4.6: MT spectral receptive fields (SRF) obtained using the WIM model. The SRF of the model mechanism (right hand panels) gradually acquires orientation relative to the spatial and temporal frequency axes as the transient neuron temporal frequency tuning (solid lines in left hand panels) changes from low-pass to band-pass. Left hand panels (a, d, g): The transience factor (f) controlling the band-pass extent of the temporal frequency function is set to 0.0, 0.2, and 0.6, respectively, in the three different panels. Middle panels (b, e, h): Spatial frequency contrast sensitivity functions for both the sustained (dashed lines) and transient (solid lines) V1 neurons. Right hand panels (c, f, i): Upper right quadrant of spatiotemporal frequency space showing contour plots of the SRFs produced by the WIM model using the temporal and spatial functions shown to the left of the plots (image taken from Perrone and Thiele (2001)).

This particular function is based on Watson and Ahumada (1985) motion model. The key parameter is the transient factor (ζ) which changes the function from low-pass or sustained ($\zeta = 0$) through to band-pass or transient ($\zeta = 1$).

The spatial frequency contrast sensitivity z_s, for each value of the stimulus spatial frequency u, is modeled using the magnitude part of the Fourier transform of a DoG functions as follows

$$z_s(u) = \sqrt{r_1^2 - 2r_1 r_2 \cos(2\pi u S) + (r_2 \cos(2\pi u S))^2 + ((1-2g) r_2 \sin(2\pi u S))^2}, \tag{4.14}$$

where $r_1 = p_1 - q_1$ and $r_2 = p_2 - q_2$, with

$$\begin{aligned} p_1 &= A_1 \exp(-xc_1(\pi u)^2) \quad ; \quad q_1 = A_2 \exp(-xs_1(\pi u)^2); \\ p_2 &= A_3 \exp(-xc_2(\pi u)^2) \quad ; \quad q_2 = A_4 \exp(-xs_2(\pi u)^2); \end{aligned} \tag{4.15}$$

This function has 10 parameters that relate the three individual differences DoG that together make up the space domain spatial receptive field. x terms are the space constants of the individual Gaussians with xc_1 and xc_2 controlling the size of the central part and xs_1, xs_2 the surrounds of the DoG. The separation between DoG is controlled by the parameter g.

2. A mechanism that responds maximally to the spatial and temporal frequencies that lie along the S-T- intersection. If T and S represent the transient and sustained neurons' contrast sensitivities, respectively, the contrast sensitivity of the WIM sensor is given by

$$M(u, \omega) = \frac{\log(T + S + \alpha)}{|\log T - \log S| + \delta}, \tag{4.16}$$

where α and δ are constants. So that:

- α helps broaden the response profile of the WIM sensor, being analogous to the background or spontaneous activity of neurons.
- δ prevents division by zero making the output less sensitive to noise. δ also controls the width of the WIM sensor and it is used to set the bandwidth of it.

4.2 MOTION MODELS

4.2.1 Classical solutions of the aperture problem

Considering the output of a single motion detector, it is not possible to know the true velocity of a moving object (aperture problem). The object seems to be moving perpendicularly to its orientation and extra information is needed to determine the real motion direction.

In the case of drifting gratings, a grating seen moving behind a circular aperture is ambiguous. But, if a second drifting grating is superimposed forming a plaid, the perceived motion is not ambiguous anymore. In order to explain the perceived motion of a plaid, three mechanisms were proposed:

1. **Intersection of constraints (IOC):** This mechanisms establishes a *contraint line* in the velocity space of all possible positions of the moving contour after an interval of time Δt. The perceived motion follows the velocity vector of the intersection in velocity space of the constraint lines of the plaid components. This mechanism was originally proposed by Adelson and Movshon (1982) to explain their results. Later, Heeger (1992) proposed a neural model to explain this procedure.

2. **Vector average (VA):** The velocity of the plaid is the vector average of the normal components of each constituent grating.

3. **Feature tracking (terminators):** In the case of plaids, the features to track where the aperture problem is solved, are the intersections. Other features are line endings and object corners.

In plaids type II[1] the direction predicted by IOC differs of the direction predicted by VA (see Figure 4.7). In the fovea, plaids type II are perceived in a direction 5° away from the IOC prediction towards the components direction. Whereas, in plaids type I[1] the IOC prediction accurately coincides with the perceived direction (Ferrera and Wilson (1990)). In peripheral vision, the perceived direction of plaids type II deviates by up to 40° from the IOC prediction (Yo and Wilson (1992)).

Also, Yo and Wilson (1992) found that for brief presentations (60ms) in the fovea, plaids type II are perceived to more in the vector average direction.

Masson and Castet (2002) studied the human perception of unikinetic plaids[1]. As in this case only one component drifts, the rules in the velocity space as IOC or VA cannot be applied to recover the perceived motion direction. This information can however be reconstructed using the motion of blobs (*feature tracking*) that are generated at the intersections between the two component gratings.

There is not an agreement about which mechanism best explain motion perception. Many models implemented the mechanisms here described adding different interactions between brain areas and within cell populations.

[1]Here some short definitions about plaid stimulus:

- **Plaid type I**: Type of plaid where the direction predicted by IOC coincides with the direction predicted by VA. These predictions are made in the velocity space considering the velocity of each drifting grating component.
- **Plaid type II**: Type of plaid where the direction predicted by IOC differs with the direction predicted by VA.
- **Unikinetic plaids**: Degenerate version of plaids type II where one component only is drifting

The following sections will describe the main bio-inspired motion models present in the literature. For a better understanding, we divided the existing bio-inspired motion models into two groups: feedforward models, which consider only the bottom-up stream from V1 to MT, and recurrent models, where feedbacks and recurrent connections are also considered. The methods here described implemented one of the mechanisms presented in this section for the aperture problem solution.

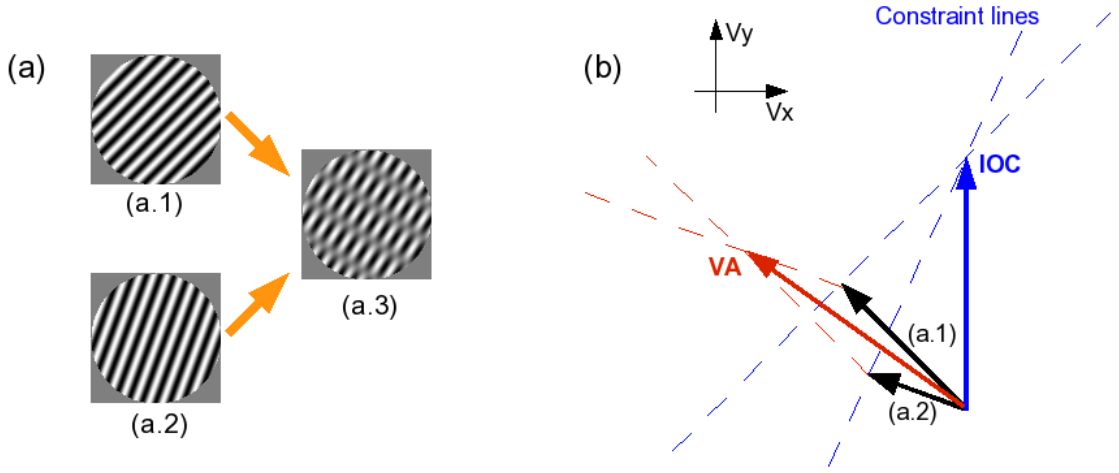


Figure 4.7: (a) Drifting gratings used to build a plaid type II, which is characterized because the direction predicted by IOC differs from the direction predicted by VA. (a.1) grating drifting in a direction of 135°, with a spatial frequency of 0.1[pixel/sec] and a temporal frequency of 6[Hz]. (a.2) grating drifting in a direction of 160°, with a spatial frequency of 0.1[pixel/sec] and a temporal frequency of 3[Hz]. (a.3) plaid obtained superimposing gratings (a.1) and (a.2). (b) Diagram showing the IOC and VA predictions in the velocity space for the motion perceived in (a.3). The motion direction perceived in (a.3) is 5°deviated from the IOC prediction towards the component directions (Ferrera and Wilson (1990)). The IOC prediction is much accurate than the VA prediction.

4.2.2 Feedforward models

Grzywacz-Yuille

In Grzywacz and Yuille (1990), the authors proposed a model for the estimation of local image velocity by cells in the visual cortex. Since the motion sensitive cells of primary visual cortex are not sensitive to local velocity, but sensitive to the direction of motion and tuned to spatiotemporal frequencies, they showed how those cells can be combined in order to estimate the local velocity of an input stimulus.

Motivated by the fact that several properties of V1 motion detector cells can be explained by motion-energy filters (see Adelson and Bergen (1985)), they introduced a method for the velocity computation from the outputs of motion-energy filters for translational motion. This velocity computation is done wiring up the filters' outputs to create a new velocity selective cell which is consistent with MT *pattern* cells.

The model proposed can be divided into two stages, equivalent with V1 and MT:

1. **Motion-energy measurement.** The motion-energy filters $N(\cdot)$ are based on the spatiotemporal oriented filter defined by

$$F(\mathbf{x}, t; \Omega, \mathbf{n}, \Omega_t, \sigma, \sigma_t) = \frac{1}{(2\pi)^{2/3}\sigma^2\sigma_l} \exp\left(-\frac{|\mathbf{x}|^2}{2\sigma^2}\right) \exp\left(-i\Omega\mathbf{n}\mathbf{x}\right) \exp\left(-\frac{t^2}{2\sigma_t^2}\right) \exp\left(-i\Omega_t t\right), \tag{4.17}$$

   where $\mathbf{x}$ and t are a spatial and temporal location in the image. $\sigma > 0$, $\sigma_t > 0$, Ω, and Ω_t are scalar parameters and $\mathbf{n} = (\cos\theta, \sin\theta)$ is a unit vector indicating the spatial orientation of the filter.

   So, the response of a directionally selective cell to an image $L(\mathbf{x}, t)$, is modeled as the nonlinear response

$$N(\mathbf{x}, t; \Omega, \mathbf{n}, \Omega_t, \sigma, \sigma_t) = |F(\mathbf{x}, t; \Omega, \mathbf{n}, \Omega_t, \sigma, \sigma_t) * L(\mathbf{x}, t)|^2, \tag{4.18}$$

   where $*$ represents convolution. This nonlinear filter is spatiotemporally tuned to a sinusoidal grating traveling in the direction $\mathbf{n}$, with spatial frequency Ω, and with temporal frequency Ω_t. σ and σ_t determining the sharpness of the tunings.

   For their analysis and for computational convenience, the authors assumed that the bandwidth of the temporal frequency tuning curve is relatively wide compared with the spatial bandwidth.

2. **Velocity estimation.** The authors proposed three theorems to find the distribution of spatiotemporal filters in order to have a maximal response for a certain velocity. The analysis is done considering that a translating image lies on the plane $\omega\mathbf{v} + \omega_t = 0$ in the frequency domain. The theorem defines that if σ and σ_t are constants, then the local maximum of $N(\mathbf{x}, t; \Omega, \mathbf{n}, \Omega_t, \sigma, \sigma_t)$ as a function of $(\Omega, \mathbf{n}, \Omega_t)$ lies on a plane $\Omega\mathbf{n} \cdot \mathbf{v} + \Omega_t = 0$ for all images that move with a constant velocity $\mathbf{v}$.

   Under some assumptions, the cells' strongest responses lie close to the plane $\Omega\mathbf{n} \cdot \mathbf{v} + \Omega_t = 0$ for an image translating with velocity $\mathbf{v}$. But, how to estimate velocity from the combination of the outputs of motion-energy cells, whose centers of receptive field lie in a single spatial location?

   Starting from the theorems proposed by Grzywacz and Yuille (1990), the authors showed three different strategies to compute velocity from the output of energy motion detectors.

   - **The ridge strategy:** This strategy proposes excitatory connections from each local motion energy filter to the velocity selective cell. Each velocity selective cell is connected to the motion energy filters most consistent with

it (see Figure 4.8). Finally a winner-take-all mechanism is used to select the strongest velocity cell.

The connection weight between the cell $(\Omega^\mu, \Omega_t^\mu, \sigma^\mu, \sigma_t^\mu)$ and the velocity selective cell tuned to the velocity $\mathbf{v}$ should be $\exp\left(-(\sigma_t^\mu)^2(\Omega_t^\mu + \Omega^\mu \cdot \mathbf{v})^2/2\right) \exp\left(-(\mathbf{v} \cdot \Omega^{u*}/k)^2\right)$, where Ω^* is orthogonal to Ω, and k is a constant parameter.

- **The estimation strategy:** This strategy computes the velocity and estimate the image's spatial characteristics simultaneously minimizing a goodness-of-fit criterion. According to the theorems proposed in Grzywacz and Yuille (1990), the response of a motion energy filter can be approached as

$$N(\mathbf{x}, t : \Omega, \Omega_t, \sigma, \sigma_t) \approx r(\mathbf{x}, t : \Omega) \exp\left(-\sigma_t^2 \frac{\Omega_t^2 + \Omega \cdot \mathbf{v}}{2}\right),$$

  where the function $r(\Omega)$ is unknown and depends on the form of the image. $r(\Omega)$ is also independent of Ω_t keeping Ω constant (see Figure 4.9). The estimation of the velocity is finally done minimizing a goodness-of-fit criterion $E(\mathbf{v}, r(\Omega))$, with respect to $\mathbf{v}$ and $r(\Omega)$. $E(\mathbf{v}, r(\Omega))$ is minimized using standard least-square fit criterion.

- **The extra information strategy:** This strategy uses the outputs of purely spatial frequency tuned cells to calculate the spatial characteristics of the image. This information can be uses to modify $r(\mathbf{\Omega})$ in the estimation strategy

Their model suggests that V1 and MT are the two stages needed for motion computation. They also claim that MT is not concerned with the aperture problem. At the period when this work was presented, all the physiological studies revealed that V1 only was capable to extract motion of one-dimension (1D) patterns, being MT by consequence in charge to solve the aperture problem. Interestingly, Grzywacz and Yuille (1990) predicted that V1 not only analyze 1D motion and it was also able to detect 2D patterns, as it was further demonstrated by, e.g., Pack and Born (2001); Sceniak et al. (2001); Jones et al. (2001). In this case, the authors claimed that MT is only in charge to pool the motion information extracted in V1.

Nowlan-Sejnowski

Nowlan and Sejnowski (1994, 1995) proposed a motion processing model to compute the two-dimensional velocities of moving objects that are occluded and transparent. Their goal, was not to have an accurate velocity representation, but instead, to segment an image into regions of coherent motion, provides an estimate of velocity in each region, and actively selects the most reliable estimates. The model uses motion-energy filters in the first stage of processing and computes, in parallel, two different

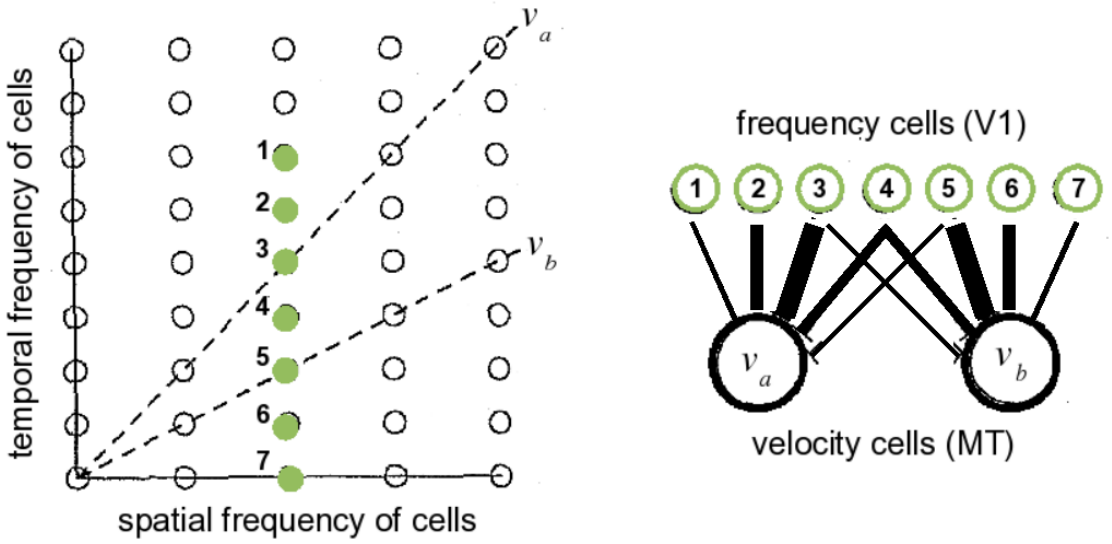


Figure 4.8: Ridge strategy for velocity computation proposed by Grzywacz and Yuille (1990). *Left:* The circles represent samplings of energy motion detectors in the frequency space. The cross sections of two velocities planes are shown (v_a and v_b) and seven motion energy cells are labelled. *Right:* Each of the seven cells have excitatory connections to the velocity cells tuned to v_a and v_b. The line width is correlated to the strength of the connection weight. Connection weights are strong if the motion energy parameters are close to the velocity plane of interest (image adapted from Grzywacz and Yuille (1990)).

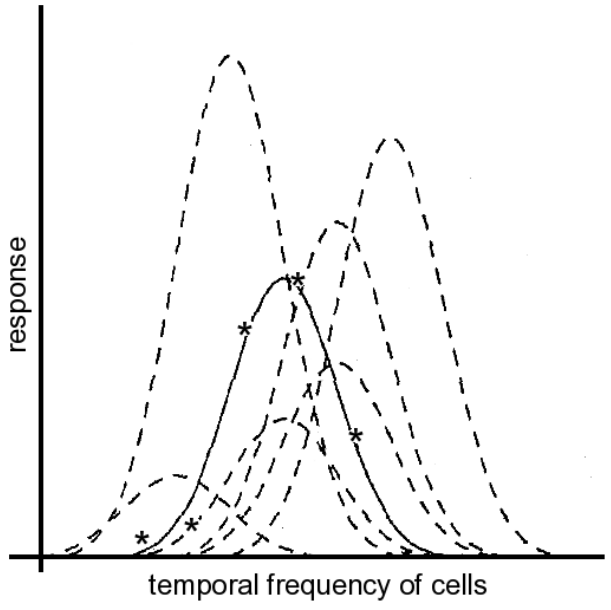


Figure 4.9: Estimation strategy for velocity computation proposed by Grzywacz and Yuille (1990). The figure shows the motion energies of a moving dot sampled by seven motion energy cells. The estimation strategy computes the image's spatial characteristics and velocity by finding the amplitude and center of the motion energy distribution. The procedure finds the best fit of the spected distribution to the data: (*) motion energies, (solid line) correct estimate, (dashed line) incorrect estimates.(image taken from Grzywacz and Yuille (1990)).

unit responses: one set of units estimates the local velocity, and the second set selects from these local estimates those that support global velocities.

The model is a feedforward cascade of locally connected networks of processing units organized into two parallel processing pathways (see Figure 4.10). The model is divided into three stages:

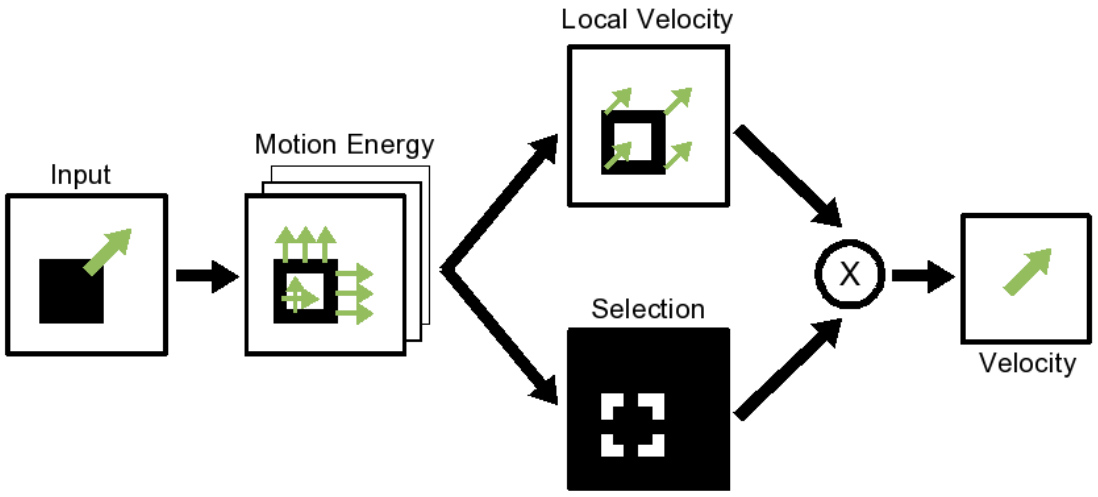


Figure 4.10: Schematic diagram of the feedforward processing model proposed by Nowlan and Sejnowski (1994) (image adapted from Nowlan and Sejnowski (1994)).

1. **Local motion energy** is extracted from all the locations in the input sequence. There are 36 motion-energy measurements for each image location (4 orientations and 9 pairs of spatial-temporal frequencies). The motion energy is extracted combining two-dimensional Gabor filters with spatial frequencies of (ω_x, ω_y) and sigma (σ_x, σ_y), together with a bandpass temporal filter of the form

$$f_k(t) = (\omega_t t)^k \exp(-\omega_t t) \left[\frac{1}{k!} - \frac{(\omega_t t)^2}{(k+2)!)} \right], \tag{4.19}$$

where ω_t is the filter center frequency and k determines the tuning width.

The outputs of motion-energy stage were organized into a grid of 49×49 receptive-field locations. For each of these receptive field locations there were 36 raw motion energy measurements. The output of the motion-energy stage is normalized using a soft-maximun normalization

$$\hat{E}_i(x,y) = \frac{\exp\left[E_i(x,y)\right]}{\sum_j \exp\left[E_j(x,y)\right]}, \tag{4.20}$$

where $E_i(x,y)$ is one of the 36 raw motion-energy measurements at location (x,y) and $\hat{E}_i(x,y)$ the corresponding normalized response, which lies between 0 and 1.

2. **Local velocities** and the validity of each velocity estimate are computed in parallel defining two different pathways: *local-velocity* pathway and *selection* pathway.

   The *local-velocity* pathway combines information from motion-energy filters tuned to different directions and spatial and temporal frequencies to find the planes of maximal motion energy in spatial and temporal frequency space, following an algorithm similar to the one proposed by Grzywacz and Yuille (1990). So, the total input to a unit tuned to velocity v_k, I'_k is defined as

$$I'_k(x,y) = \sum_{\omega_x,\omega_y,\omega_t} w_{k,\omega_x,\omega_y,\omega_t} \hat{E}_{k,\omega_x,\omega_y,\omega_t}(x,y), \tag{4.21}$$

   where the weights $w_{k,\omega_x,\omega_y,\omega_t}$ are inversely proportional to the distance between the plane defined by velocity v_k and the center frequency of each motion-energy unit.

   Velocity units receive inputs from all motion-energy units, with directional preferences within $\pm 90^\circ$ from the preferred direction of the velocity unit. The weights between the motion-energy units and the velocity-tuned units were trained to optimize a global measure of performance of the model. All the pools of velocity-tuned units shared a common set of weights. The velocity at a receptive-field location was represented by the relative strengths of the input to each velocity unit, this is done through the soft-maximum nonlinearity:

$$I_k(x,y) = \frac{\exp\left[I'_k(x,y)\right]}{\sum_j \exp\left[I'_j(x,y)\right]}, \tag{4.22}$$

   where $I_k(x,y)$ is the final state of the unit representing velocity v_k and $I'_k(x,y)$ is the initial state of the unit.

   The *selection* pathway estimates the local validity of each velocity estimate calculating a support index for each of them. The support $S'_k(x,y)$ assigned to each location (x,y) for the velocity hypothesis v_k is computed as

$$S'_k(x,y) = \sum_{\omega_x,\omega_y,\omega_t} \hat{E}_{\omega_x,\omega_y,\omega_t}(x,y), \tag{4.23}$$

   where the weights $w_{k,\omega_x,\omega_y,\omega_t}$ were initialized randomly and their final values are determined by an optimization procedure.

   The constraint on the total amount of support for each hypothesis was enforced by use of global competition among all the units in the each selection layer, which was implemented with a soft-maximum nonlinearity:

$$S_k(x,y) = \frac{\exp\left[S'_k(x,y)\right]}{\sum_{x',y'} \exp\left[S'_k(x',y')\right]}, \tag{4.24}$$

   where $S'_k(x,y)$ is the net input to a selection unit in layer k and $S_k(x,y)$ is the output state of that unit.

3. **Global estimates** of the velocities of objects within the visual scene are formed by integration across subsets of the local-velocity estimates according to the relative confidence values assigned by the selection pathway. The global evidence for a visual target moving at a particular velocity $V_k(t)$ is computed as a sum, over the product of the outputs of local velocity and selection pathways:

$$V_k(t) = \sum_{x,y} I_k(x,y,t) S_k(x,y,t), \tag{4.25}$$

where $I_k(x,y,t)$ is the local evidence for velocity k computed by the velocity pathway from region (x,y) at time t and $S_k(x,y,t)$ is the weight assigned by the selection pathway to that region. The weights were adjusted to optimize a measure of the overall performance of the model.

The main difficulties to implement the model proposed by Nowlan and Sejnowski (1994, 1995) comes from the estimation of the connection weights $w_{k,\omega_x,\omega_y,\omega_t}$ (see (4.22)*). Their method is to run an optimization algorithm in the training stage. After the weights are fixed, the system can be applied to similar input stimuli. One limitation is that the model does not deal with temporal integration of motion. As a consequence, for example, their model does not detect non-Fourier (second-order) motion stimuli.*

Simoncelli-Heeger

Simoncelli and Heeger (1998) proposed a two-stage physiological model for local image velocity representation in visual areas V1 and MT. These two areas are the two primary stages of the model. In the two stages the treatment of the signal is the same: a weighted sum of input values followed by rectification, squaring and response normalization.

Thanks to the elements previously described in Chapter 3 and Section 4.1, we can briefly describe each part of the model as follows:

1. **V1 simple cells:** They are modeled by linear receptive fields (energy motion detectors selective for spatiotemporal orientation), followed by a half-quaring rectification and a divisive normalization. The divisive normalization attempt to account some simple cell nonlinearities, such as, contrast saturation and cross-orientation inhibition. The divisive normalization is performed dividing the response of each neuron by a quantity proportional to the summed activity of a pool of neurons inside a cortical neighborhood. The cortical neighborhood is formed with neurons tuned to a the full range of orientation, direction and spatiotemporal frequency.

2. **V1 complex cells:** They are modeled to have a response relatively independent of the precise stimulus position within the receptive field. This is attained

computing V1 complex cell response as a weighted sum of V1 simple cells distributed over a local spatial region having the same spatiotemporal orientation and phase. As V1 motion detectors are modeled as energy-filters, V1 complex cells are not selective to stimulus velocity. They are only selective to the component of velocity orthogonal to their preferred spatial orientation.

3. **MT pattern cells:** They are modeled in order to create a velocity detector with a IOC velocity detection mechanisms. The IOC-like behavior is obtained summing the responses of a particular set of V1 neurons. The spatiotemporal frequency bands of the V1 neurons summed are bisected by the plane of the translating twodimensional pattern (see Figure 4.11). The summation is over both temporal and spatial frequency. The MT cell obtained have broader spatial frequency bandwidths than the V1 neurons. In addition to the summation over spatiotemporal frequency, each MT neurons sums the responses of V1 neurons with receptive field positions in a local spatial neighborhood. Finally, the MT responses are half-squared and normalized, as in V1 stage.

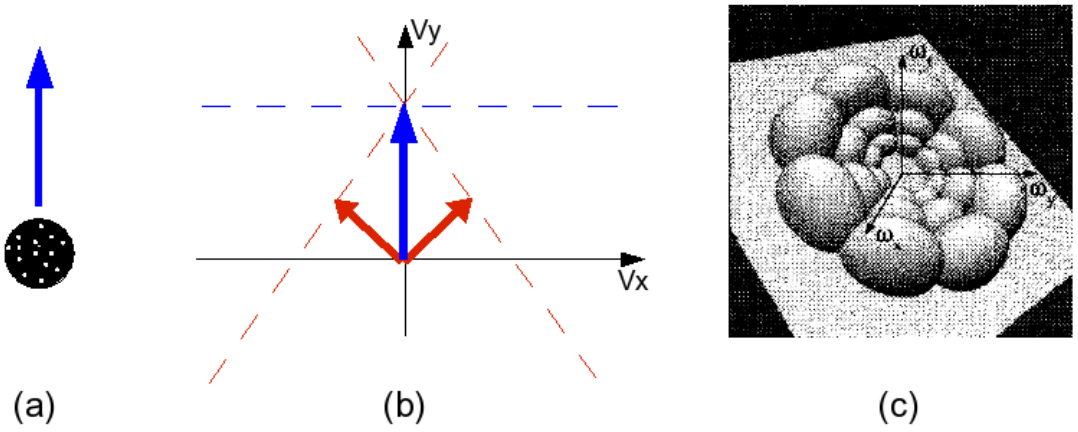


Figure 4.11: Construction of MT pattern cell velocity selectivity using a combination of V1 complex cells. (a) Random dot field stimulus drifting upwards. (b) Intersection of constraints (IOC) construction for the stimulus shown in (a). Red arrows correspond to the normal component of velocity for two possible pair of V1 complex cells satisfying the motion direction (blue arrow). (c) Selectivity of V1 neurons tuned for four orientations and three spatial scales, each consistent with a common velocity, the velocity tuned for the MT cell. These neurons are summed with a positive weight to yield a MT neuron sensitive to this velocity (image adapted from Simoncelli and Heeger (1998)).

The MT neurons modeled as isolated entities cannot encode stimulus:

- A neuron's response depends on stimulus contrast and spatial pattern.
- Even for a fixed contrast and spatial pattern, there are family of velocities evoking the same response (arranges in concentric contours around the preferred velocity).

The representation of velocity is implicitly encoded in the simultaneous response of the population of MT neurons. The response of MT population must be interpreted as discrete samples of a continuous two-dimensional response distribution (velocity space).

The simulations performed by Simoncelli and Heeger (1998) showed that the model is consistent with a a variety of physiological data.

The major concern with this model is the lack of realistic temporal dynamics, where the outputs of each neuron correspond to steady-state firing rates. Another limitation, also related with the time treatment, is the choice of a Gaussian function for the temporal part of the V1 linear motion detector. This choice is computationally convenient, but it has three main drawbacks. First, Gaussian derivatives at different scales produce an uneven tiling of the Fourier domain. Second, the resulting spatial and temporal frequency tuning curves are not separable; In fact, the spatiotemporal frequency tuning curves are polar-separable which is inconsistent with V1 physiology. Third, Gaussian derivative receptive fields are non-causal. This problem can be solved introducing a time delay, but a more appropriate solution is to use recursive temporal derivative filters.

Giese-Poggio

Giese and Poggio (2003) studied the contribution of both visual information pathways: form and motion, in a real application as biological motion recognition. They proposed a hierarchical feedforward neural model, where the size of the receptive fields is gradually increasing. Figure 4.12 shows a diagram summarizing their model.

The form pathway is modeled as

- Local orientation detectors modeled by Gabor filters (V1). They used eight different orientations and two spatial scales differing by factor 2. Cells are placed in a equidistant grid and their receptive field sizes are according to real observations in monkey V1 simple cells.
- Position-and-scale-invariant bar detectors (V1,V4). The invariance (within a certain range) is obtained pooling the responses of neurons with similar preferred orientation inside a neighborhood. The total activation is done considering the max operator between the cells at different scale. A linear threshold is used to feed this response to the next layer.
- View-tuned neurons (IT, STS, FA[2]). This stage is formed by a recurrent neural network where the neurons are previously trained for a certain action (view-

[2]Some basic definitions: IT: infotemporal cortex, STS: superior temporal sulcus, FA: fusiform face area, F5: area in monkey premotor cortex.

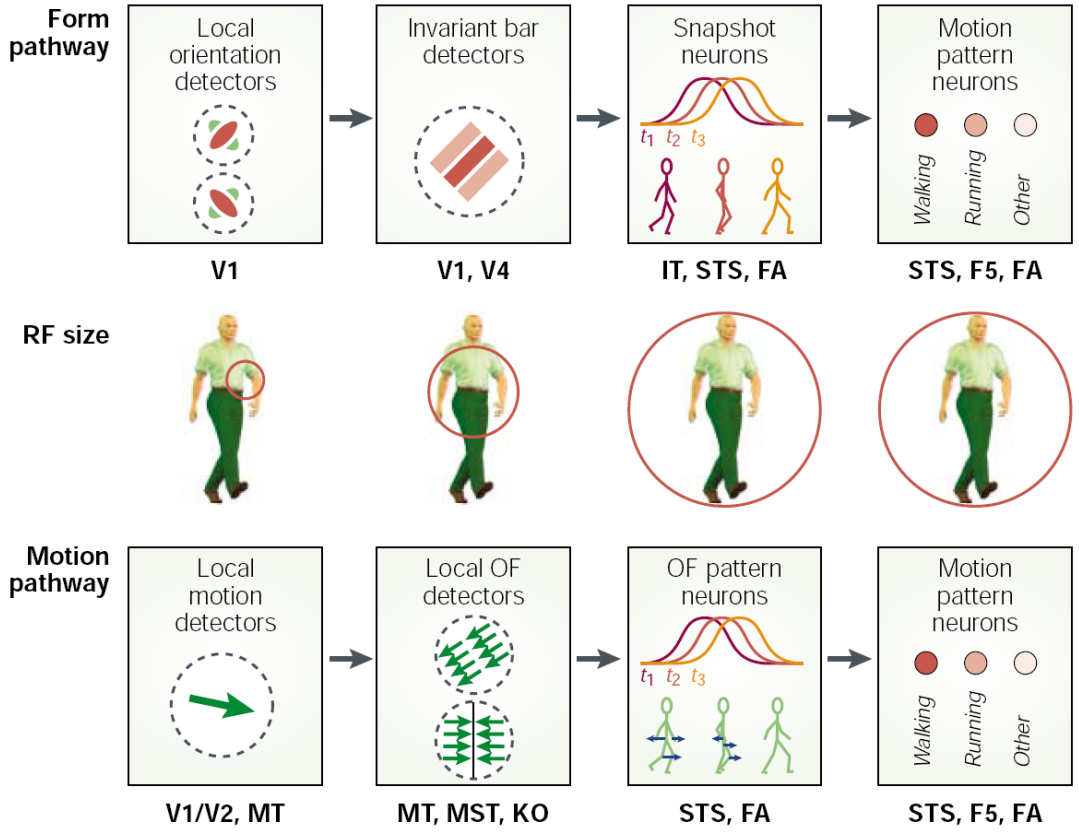


Figure 4.12: Diagram summarizing the hierarchical neural model proposed by Giese and Poggio (2003). The diagram is an overview of the model showing the two pathways for the processing of form and motion. The middle row show the size of the receptive fields of each stage compared to typical stimulus (image taken from Giese and Poggio (2003)).

tuned neurons). The output of the form pathway is the sum over all the view-tuned units representing the same biological motion pattern. Snapshots neurons are selective, for instance, for body shapes. They have large receptive fields and show substantial position and scale invariance. The snapshots neurons were modeled by *Gaussian radial basis function* (RBF), which centers are adjusted during training.

- Motion pattern neurons (STS, F5, FA[2]). These neurons are modeled by a dynamic equation which temporally smooth and average the activity of all snapshot neurons that contribute the encoding of the same movement pattern (for details, see Chapter 6).

The modeling of the motion pathway, which concerns this thesis, is modeled as follows

- Local motion detectors corresponding to V1 direction-selective neurons and component motion-selective neurons in area MT (see Section 3.2.4). They did not model the extraction of the local motion energy in detail, instead they calculated the optic flow fields directly from the stick figure model that was animated

using two-dimensional tracking data from video sequences.

- This stage is formed by neurons with larger receptive fields that analyze the local structure of the optic-flow fields induced by motion stimulus. They proposed two types of optic-flow detectors

  1. Neurons selective for translation flow, which corresponds to neurons in area MT, with low or bandpass tuning respect to speed. They included four populations of neurons with four preferred directions and with receptive field sizes similar to those found in MT neurons.
  2. Neurons selective for motion edges (horizontal and vertical), which corresponds to neurons found in MT, MSTd and MSTl. Their outputs are built combining the responses of two adjacent subfields with opposite direction preferences in a multiplicative way. These neurons also present scale invariance. The scale invariance is obtained pooling the signals inside the neuron receptive field using a maximun operator.

- Optic-flow pattern neurons. Equivalent to snapshot neurons of form pathway modeling, whose existence is a prediction of the model. These neurons, also modeled by RBF, are selective to complex patterns of optic flow and their parameters are obtained after a training procedure. The assumed that this type of neurons can be found in STS, FA[2] and maybe MST.

- Motion pattern neurons. Analogously to motion pathway modeling, the outputs of optic-flow pattern neurons are summed and temporally smoothed. They assumed that this type of neurons can be found in monkey areas STS, FA and F5[2].

The aspects of this model concerning biological motion recognition will be talked at length in Chapter 6.

This model was implemented following a specific goal: to summarize the mechanisms involved in the processing of biological motion. The authors implemented a feedforward model processing the form and motion information independenly. Their model was only tested with biological motion stimuli. About the limitations of the model we could cite: no inclusion of attentional mechanismsm or eye movements; no feedback connections or interactions between the dorsal and ventral pathways, no biological plausibility for the motion detectors units.

4.2.3 Recurrent models

Unlike feedforward models, recurrent models include all the models with feedbacks and recurrent connections. In most cases, the feedbacks will emphasize the activation of a certain population of V1 neurons.

Wilson, Ferrera and Yo

The authors proposed in Wilson et al. (1992) a motion processing model where MT neurons receive two different inputs coming from two different parallel motion pathways. The MT units receiving a weighted sum of the outputs of these two pathways start a competitive feedback mechanism which extracts the maximum response. A schematic diagram of this two motion pathways is shown in Figure 4.13.

The model of Wilson et al. (1992) has two parallel motion processing pathways:

1. The simple one (Fourier motion, Figure 4.13 *Left*) consists of orientation-selective filtering followed by motion-energy extraction and a contrast gain control stage. A contrast gain stage is here necessary to introduce a gain-control operation to minimize the effects of component contrast variations (contrast saturation). The contrast normalization is calculated dividing the output of the motion-energy stage by the output of the orientation-selective filters.

2. The second pathway (non-Fourier motion, Figure 4.13 *Right*) also provides inputs to MT neurons and it extracts the motion of texture boundaries. In order to detect texture boundaries, an additional processing is needed before the motion-energy extraction and contrast gain control stage. The additional processing employs: filtering of the input image, response squaring (or rectification) and a second filtering stage with a lower frequency. This mechanism has been previously used to extract the location of texture boundaries or discontinuities from images (see, e.g., Bergen and Landy (1991); Landy and Bergen (1991)). This additional processing in the texture boundary motion pathway is suggestive of existing processing in area V2.

The final stage of the model combines the inputs coming from the two motion pathways (the simple motion energy pathway and the texture boundary motion pathway) using a cosine weighted function. The units of this stage compute the direction of pattern motion. If R_i is the response of the ith input unit with preferred direction θ_i, and θ_p is the preferred direction of a given pattern unit, the input to this pattern unit, E_p, is given by

$$E_p = \sum_{i=2}^{N} R_i \cos(\theta_p - \theta_i), \tag{4.26}$$

where N is the number of direction of motion of the Fourier and non-Fourier pathways.

Following the computation of E_p, there is a stage of competitive feedback mechanism within MT neurons. This mechanism is designed to extract the response of the most strongly stimulated pattern unit. Each MT pattern unit sends inhibitory signals to MT pattern units tuned with a relative difference of angle $\Delta\theta$ of $45^\circ \leq |\Delta\theta| < 90^\circ$. Finally, after the recurrent inhibition stage, the final predicted direction of motion is obtained by parabolic interpolation. The effect of the inhibitory feedback is to extract

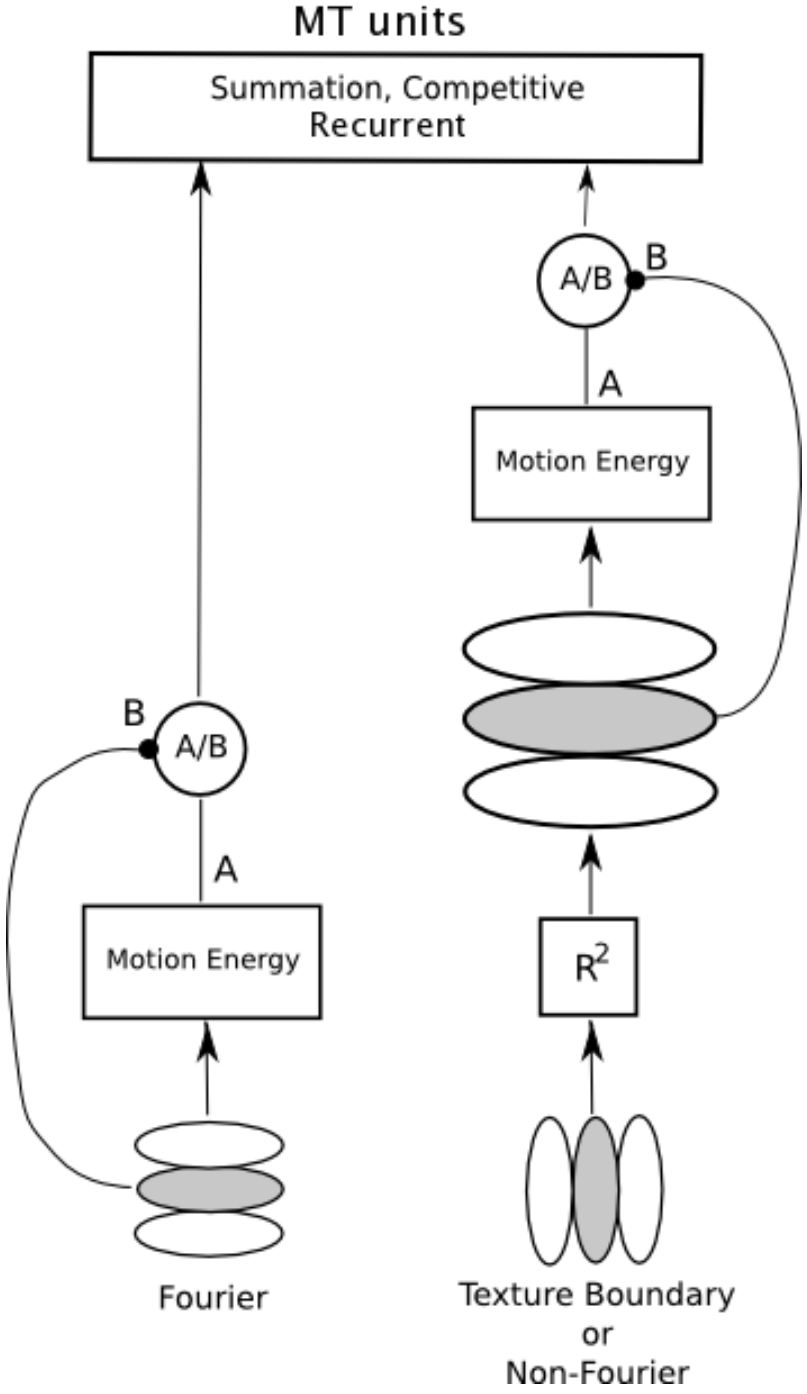


Figure 4.13: Schematic of the two-dimensional motion processing pathways proposed by Wilson et al. (1992). The simple Fourier pathway on the left starts with orientation-selective filtering followed by a motion energy computation. The response of the oriented filters also provides a feedforward gain-control signal that divides the motion energy output (circle A/B). The texture boundary pathway, or non-Fourier pathway, on the right computes the motion energy and gain control after an oriented filtering, squaring, and a second stage of filtering at a different orientation and lower spatial frequency. The competitive inhibition at the final stage (top) extracts the maximum response (image taken from Wilson et al. (1992)).

the maximum plus nearest neighbors of the pattern unit responses. The direction of motion obtained through the computation of the maximum of the cosine-weighted sum is equivalent to the VA direction.

This model was only compared with real data measured on plaids composed of two grating components, accurately predicting their perceived direction. The model calculates the vector average solution only for a few periods of time. The non-Fourier pathway shifts the calculated direction away from VA towards the IOC direction.

Grossberg et al.

The motion Boundary Control System (BCS) model proposed by Chey et al. (1997) and Grossberg et al. (2001) attempts to find the solution to the global aperture problem by showing how information from feature tracking points, where unambiguous motion can be computed, can be propagate to ambiguous motion direction points and resolve the real motion direction.

The model is summarized in Figure 4.14. The first few stages of the model use transient cells that feed a multiscale short-range motion filter whose larger scales selectively process higher speeds as a result of the combined action of self-similar thresholds and competition.

Getting more into details, each stage of the model can be described as

- **Level 1: Input.** The FACADE model (Grossberg and Mingolla (1985)) is used to extract the T-junctions of the input stimulus without using a T-junction detector. It uses circuits that include oriented bipole cells modeling existing cells in V2. The output of this first level, is a binarized sequence containing only the T-junctions of the input stimulus.

- **Level 2: Transient cells.** The second stage is formed by unidirectional transient cells, directional interneurons and directional transient cells. Unidirectional cells respond to motion (changes in luminosity) independently of its direction. By the contrary, directionally selective neurons highly respond to motion on their preferred direction while little response is evoked to motion in the reverse direction. These three types of cells are connected between them following these three principles:

  (a) Directional selectivity is the result of asymmetric inhibition along the preferred direction of the cell.

  (b) Inhibition in the null direction is spatially offset from excitation.

  (c) Inhibition arrives before, and hences prevents, excitation in the null direction.

- **Level 3: Short-range filter.** The mechanism proposed in Layer 2 presents two inconveniences due to the local null-direction inhibition process, which in part

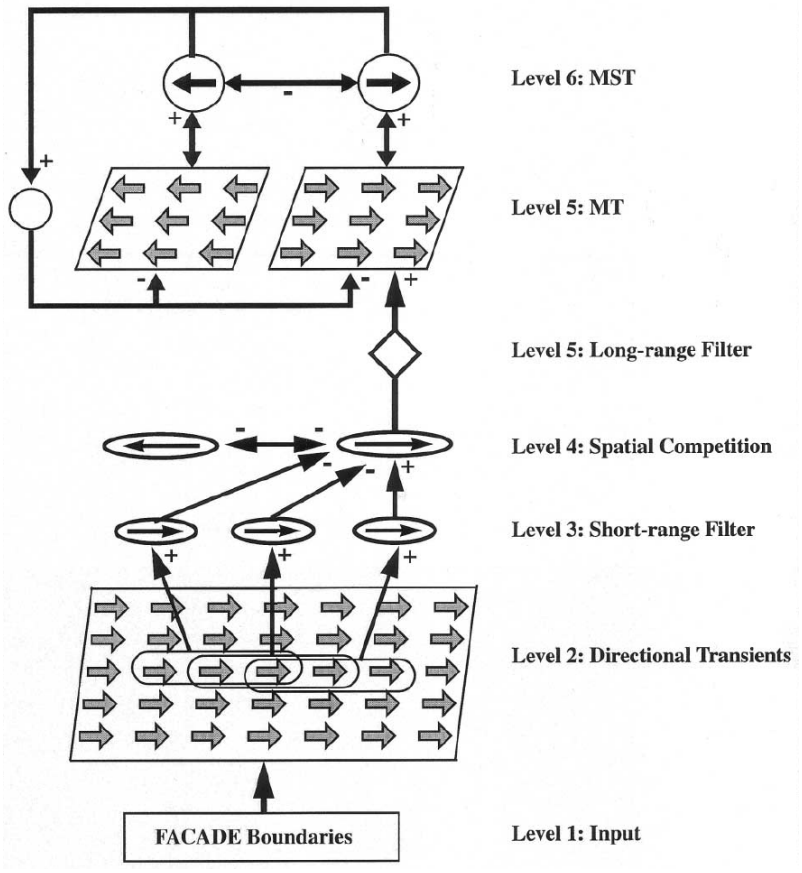


Figure 4.14: Diagram summarizing the five stages of the model proposed by Chey et al. (1997). (image taken from Grossberg et al. (2001)).

does not selectively activates the correct direction. In this layer, the directional transient cells are spatially and temporally averaged by a short-range filter. The short-range filter only accumulates evidence from directional transient cells of similar directional preference within a spatially anisotropic region that is oriented along the preferred direction of the cell. The short-range filtering is done at different spatial scales, each of them corresponding to a different speed range.

- **Level 4: spatial competition and opponent direction inhibition.** In order to enhance the amplitude of feature-tracking signals compared to ambiguous signals, a spatial competition is proposed in this layer. The spatial competition is done among cells of the same spatial scale preferring the same motion direction. This model stage also uses opponent inhibition between cells tuned to opposite directions, avoiding of this way a simultaneous enhancement.

- **Level 5 and 6: Long-rage filter, directional grouping, and attentional priming.** As it is shown in Figure 4.14, Layers 5 and 6 are linked by a feedback network. Layer 5 attempts to models MT cells using a spatially long-range filter. Similarly to Layer 3, but in a larger spatial region, the long-range filter pools signals with similar directional preference, opposite contrast polarity, and multiple orientations. This procedure makes MT neurons to react as true directional cells.

- **Layer 6: MST modeling.** The signals from MT (Layer 5) activates MST neurons, which interact via winner-take-all competition across directions. The winning direction is then fed back down to MT through an attentional priming pathway which influences a region surrounding the spatial location of the MST cell. This attention mechanism non-specifically inhibits the activation of MT neurons. In the case of the winning direction, excitation cancels inhibition, so the winning direction survives the top-down matching process.

Numerous extensions not treated in this chapter have been proposed by the authors, such as, the manipulation of extrinsic and intrinsic junctions (Grossberg et al. (2001)), form-modulated motion diffusion thanks to the FORMOTION model that can explain the changes in perception due to contextual changes (Berzhanskaya et al. (2007)), etc.

Unlike most of the models we revisited, this model propose a solution for the aperture problem diffusing in time the unambiguous motion information of feature points. Their model explains a large number of psychophysical observations but it is highly complex. The mathematical analysis of the full model is impossible and many parameters must be tuned, which makes it hard to run or propose predictions. Morever, because its high computational cost, only small binary inputs can be treated with only

4 motion directions. Another problem with such complex architectures is that we do not know which part of the system is the most critical to explain a given observation.

Bayerl-Neumann

Attempting to solve the aperture problem using contextual information, Bayerl and Neumann (2004) proposed a powerful feedforward and feedback V1-MT model. Further extensions of this preliminary model have been proposed in Bayerl and Neumann (2005, 2007). The recent implementations propose solutions to deal with transparent motion and to eliminate extrinsic junctions using inhibitory "T" junctions. In this section, we will only describe the basis of the preliminary model presented in Bayerl and Neumann (2004) (see Figure 4.15).

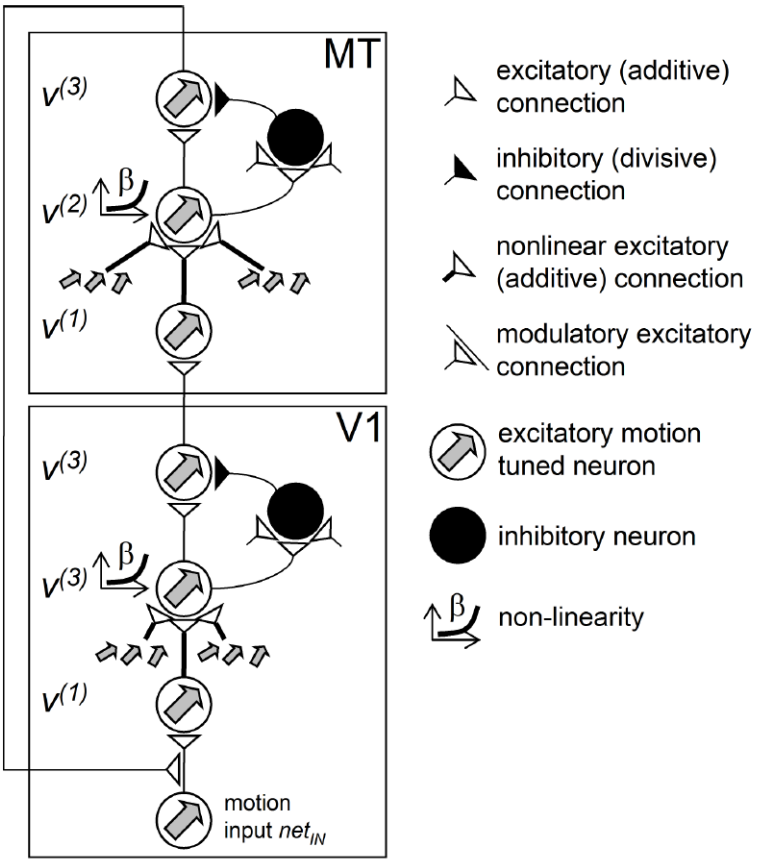


Figure 4.15: Overview of V1 and MT modeling proposed by Bayerl and Neumann (2004). The architecture shows the dynamic interaction modeled between the two visual areas: modulatory feedbacks, feedforward interaction, and lateral shunting inhibition (image taken from Bayerl (2005)).

The feedback mechanism modeled in Bayerl and Neumann (2004) between MT and V1 areas triggers a filling-in process along boundaries to solve the aperture problem.

V1 motion detectors are implemented as Elaborated Reichardt Detectors (see Section 4.1.3), where the spatial filtering is performed using spatial receptive fields se-

lective to static-oriented contrast at a fixed spatial frequency and independent of contrast polarity between two frames. The directionally-selective cells come from the pooling over all orientation-selective cells at different time steps. To keep the input as clean as possible, they modeled V1 cells as motion-sensitive cells independent of contrast orientations. They claimed that this property does not constrain the capability of the model to solve the aperture problem. The model can be adapted in order to include some important V1 cell characterization, such as, spatiotemporal frequency tuning, contrast polarity and orientation tuning.

Both, V1 and MT areas are modeled with similar architectures performing the following three stages:

1. Feedback modulation:

$$\partial_t v^{(1)} = -v^{(1)} + net_{IN} \cdot (1 + C \cdot net_{FB}) \,. \tag{4.27}$$

2. Feedforward integration:

$$\partial_t v^{(2)} = -v^{(2)} + (v^{(1)})^2 * G_{\sigma_1}^{(\mathbf{x},space)} * G_{\sigma_2}^{(\Delta\mathbf{x},velocity)}. \tag{4.28}$$

3. Lateral shunting:

$$\partial_t v^{(3)} = -0.01 v^{(3)} + v^{(2)} - \left(\frac{1}{2n} + v^{(3)}\right) \cdot \sum_{\Delta\mathbf{x}} v^{(2)}, \tag{4.29}$$

where n denotes the number of cells tuned to different velocities at any specific location, net_{IN} is the input of the model area (e.g., in V1 is the output of the motion detectors), net_{FB} is the feedback signal (e.g., again in V1 the output of MT model), $*$ denotes convolution and G_{σ_1} and G_{σ_2} are Gaussian kernels in space and velocity domain, respectively.

The main differences between V1 and MT are the spatial size of receptive fields (V1:MT, 5:1) and the mechanisms proposed within and between each areas.

The authors are focused in the feedback mechanism. Feedback mechanism works as a predictor that enhances those signals in the lower (V1) area that are compatible with respect to feature specificity. In other words, only compatible patterns get emphasized in the lower area and no activity is produced where no signal is provided by the input. Their modulatory feedback mechanism (4.27) can be compared with the IOC mechanism of motion integration, since the input signal constrains the feedback signal and the intersection of both is emphasized.

The input to the integration stage (4.28), is squared and processed by cells with isotropic spatial and isotropic directional Gaussian receptive fields. Finally, cell activities are normalized using lateral shunting inhibition (4.29). The sum of cell activities sensitive to any velocity at a specific location normalizes the total energy. By this mechanism, unambiguous signals get emphasized, while ambiguous signals lead to a flat population response.

Starting from the work of Bayerl and Neumann (2004), Tlapale et al. (2008) recently proposed a form-motion integration model. The model of Tlapale et al. (2008) solves the motion integration problem with a spatiotemporal diffusion which is modulated by luminance. They proposed an anisotropic integration model where motion diffusion is gated by luminance distribution in the image. This model explains several psychophysical experiments.

The approach presented by Bayerl and Neumann (2004) models the functionality of the main mechanism in the primate visual system. Their model was successfully tested with natural sequences showing comparative results with the ones obtained in the computer vision community. This is something new compared to the previous bio-inspired motion models, where no natural images were tested.

Based on a classical linear/non-linear model with rectification and inhibitive division as in Simoncelli and Heeger (1998), the model proposed by Bayerl and Neumann (2004) manages spatial diffusion through a feedback loop. Since the diffusion mechanism is relatively simple compared to existing ones (see for instance, Grossberg et al. (2001), it can be applied on standard sequences used in the computer vision community.

CHAPTER 5

V1-MT: CORE ARCHITECTURE

"We cannot think about what a feedback interaction could do if we do not first explore the limitations of a feedforward model"
– Simon Thorpe (GDR-vision meeting 2008)

Contents

OVERVIEW

This chapter presents a feedforward V1-MT core architecture that we are going to exploit in the rest of this thesis. This core model will be mostly classical in its conception (inspired from biology and based on existing models), but we will investigate how to extend it depending on the applications.

Starting from some V1-MT characteristics shown in Chapter 3 and the modelization elements shown in Chapter 4, can we build a simple V1-MT model to process visual motion information that can be usefully applied into a real application? Here in this chapter we attempt a V1-MT neuron models to extract motion information from an input video that will be later used for a specific task, as e.g., human action recognition[1].

Figure 5.1 shows the feedforward V1-MT core architecture proposed in this thesis to process input video streams. The model is basically formed by two layers of cortical neurons: V1 layer and MT layer. These layers are created using the V1 and MT neurons proposed in this thesis.

In this chapter we present general characteristics of the V1 and MT neurons proposed. For V1, we explain how the motion detectors are defined, while for MT, we describe a general framework that will be further talked at length in Chapters 7 and 8.

Finally, we show how those V1 and MT neural entities are combined in order to process a wide field motion. Directionally-selective filters modeling V1 simple and complex cells are applied over each frame of the video input video (see Figure 5.1 (b)). V1 neurons output feed the subsequent MT layer, which integrates the information in space and time (see Figure 5.1 (c)).

Keywords: V1, MT, simple cell, complex cell, surround interactions, energy filters, spatiotemporal filtering.

Contributions of this chapter

1. Proposition of a bio-inspired feedforward V1-MT core architecture. This core architecture will be further extended for a specific application.

2. Frequency-based analysis of the V1 motion detectors. This analysis will show the relationship between different parameters and the spatiotemporal frequency tuning.

[1]Further, in Chapters 7,8 and 9 the model here proposed is used and extended for specific applications: human action recognition and the study of the role of V1 surround suppression in the solution of the aperture problem. We show how an analog (Chapter 7) and a subsequent spiking (Chapter 8) implementation can convey successfully recognition.

3. Modelization of several MT center-surround interactions and different surround geometries.

Organization of this chapter:

Section 5.1 shows the implementation of the V1 motion detectors. Section 5.2 shows the MT basic entities. Finally, Section 5.3 describes the connectivity between V1 and MT layers.

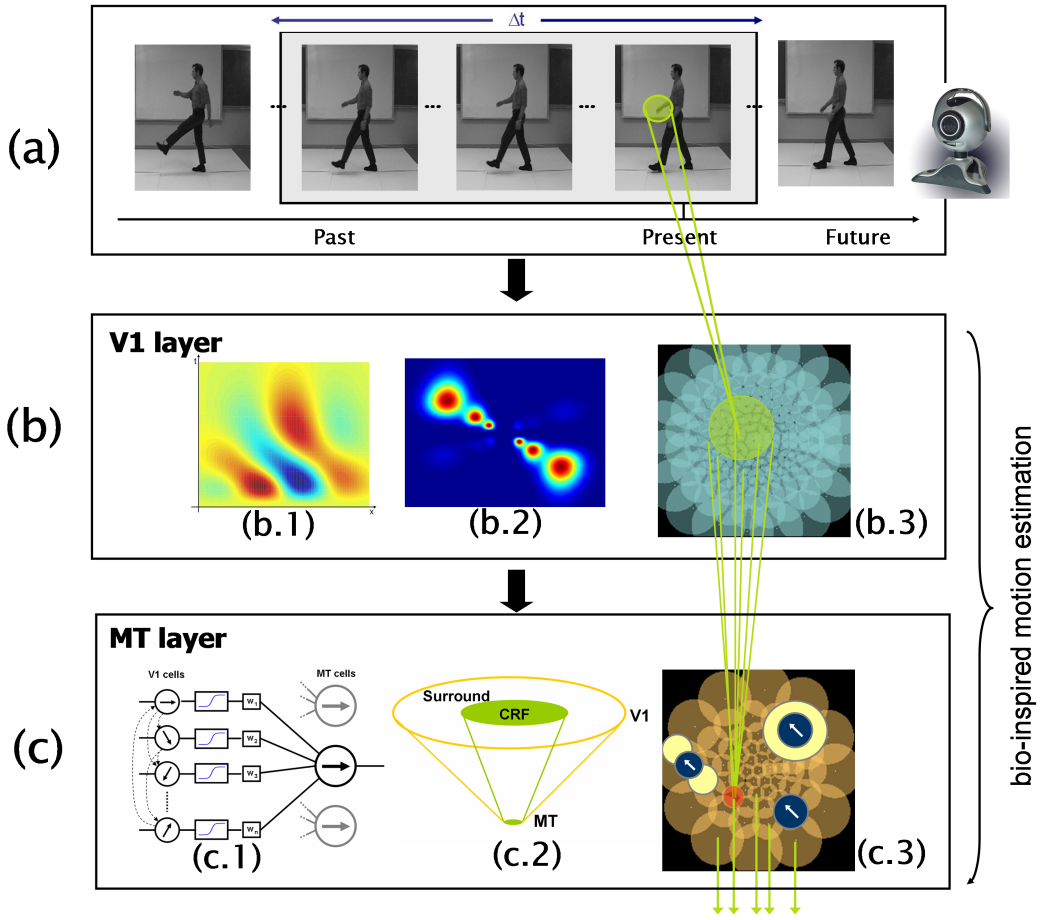


Figure 5.1: Block diagram of the feedforward V1-MT core architecture. (a) Input is a real video sequence, which is preprocessed in order to have contrast normalization and centered moving stimulus. In practice, we will consider a sliding temporal window of length Δt. (b) V1 layer: Directionally-selective filters are applied over each frame of the input sequence in a log-polar distribution grid obtaining the activity of each V1 cell. (c) MT layer: V1 outputs feed the MT cells which integrate the information in space and time.

5.1 V1: THE MOTION DETECTORS IMPLEMENTED

In Grzywacz and Yuille (1990), the authors showed that several properties of simple/complex cells in V1 can be described by energy filters and in particular, by Gabor filters. The individual energy filters are not velocity tuned, however it is possible to use a combination of them in order to have a velocity estimation.

More recently, Mante and Carandini (2005) showed which properties of V1 neurons can be explained using an energy model. In particular, Mante and Carandini (2005) demonstrated that an energy model of V1 exhibits similar behaviors to those measured with optical imaging (Basole et al. (2003)).

In this thesis, V1 neurons are modeled as energy motion detectors. The energy motion detector models a V1 complex cell which is built as a nonlinear combination of V1 simple cells.

5.1.1 V1 simple cells

Simple cells are characterized with linear receptive fields where the neuron response is a weighted linear combination of the input stimulus inside its receptive field. By combining two simple cells in a linear manner it is possible to get directionally-selective neurons, that is, simple cells selective for stimulus orientation and spatiotemporal frequency (see Figure 5.3 (a)).

As previously mentioned in Section 3.1, the direction-selectivity (DS) of a neuron refers to the property of that neuron to respond selectively to the direction of the motion of a stimulus. The way to model this selectivity is to choose receptive fields oriented in space and time, as it was described in Section 4.1.3.

Given an input stimulus $L(\mathbf{x}, t)$, the response of a spatiotemporal oriented V1 simple cell $F^*(\mathbf{x}, t)$ is obtained by the convolution

$$L(\mathbf{x}, t) * F^*(\mathbf{x}, t), \tag{5.1}$$

where $F^*(\mathbf{x}, t)$ can be defined by one of the following filters

$$\begin{aligned} F^a(\mathbf{x}, t) &= F^{odd}(\mathbf{x})H_{fast}(t) - F^{even}(\mathbf{x})H_{slow}(t), \\ F^b(\mathbf{x}, t) &= F^{odd}(\mathbf{x})H_{slow}(t) + F^{even}(\mathbf{x})H_{fast}(t), \end{aligned} \tag{5.2}$$

which are spatially located at $\mathbf{x} = (x, y)$.

The spatial and temporal parts of (5.2) are defined as follows:

- The spatial parts $F^{odd}(\mathbf{x})$ and $F^{even}(\mathbf{x})$ of each conforming simple cell derive from Gabor function $G(\mathbf{x})$ defined by

$$G(\mathbf{x}) = \exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\sin(\eta\mathbf{k}\mathbf{x}), \tag{5.3}$$

where $\mathbf{k} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$ and $\eta = 2\pi f$. f is the spatial frequency of the Gabor function, σ its standard deviation and θ its spatial orientation.

More precisely, starting from (5.3), $F^{odd}(\mathbf{x})$ and $F^{even}(\mathbf{x})$ are then defined by

$$F_\theta^{odd}(\mathbf{x}) = \frac{\partial G_\theta(\mathbf{x})}{\partial x} \tag{5.4}$$
$$= \exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\left[\eta\cos(\eta\mathbf{kx}) - \frac{\mathbf{kx}}{\sigma^2}\sin(\eta\mathbf{kx})\right],$$
$$F_\theta^{even}(\mathbf{x}) = \frac{\partial^2 G_\theta(\mathbf{x})}{\partial x^2} \tag{5.5}$$
$$= \exp\left(-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right)\left[\left(\frac{\mathbf{k}^2\mathbf{x}^2}{\sigma^4} - \eta^2 - \frac{1}{\sigma^2}\right)\sin(\eta\mathbf{kx}) - \frac{2\eta\mathbf{kx}}{\sigma^2}\cos(\eta\mathbf{kx})\right],$$

- The temporal contributions $H_{fast}(t)$ and $H_{slow}(t)$ derive from the Gamma functions $T_{\eta,\tau}(t)$ defined as

$$T_{\eta,\tau}(t) = \frac{t^\eta}{\tau^{\eta+1}\eta!}\exp\left(-\frac{t}{\tau}\right), \tag{5.6}$$

where τ specifies the time decay of (5.6) and η its order.

More precisely, starting now from (5.6), $H_{fast}(t)$ and $H_{slow}(t)$ are defined subtracting two Gamma functions with a difference of two in their respective orders obtaining

$$H_{fast}(t) = T_{3,\tau}(t) - T_{5,\tau}(t),$$
$$H_{slow}(t) = T_{5,\tau}(t) - T_{7,\tau}(t), \tag{5.7}$$

The biphasic shape of $H_{fast}(t)$ and $H_{slow}(t)$ could be a consequence of the combination of cells of M and P pathways (De Valois et al. (2000); Saul et al. (2005)) or be related to the delayed inhibitions in the retina and LGN (Conway and Livingstone (2003)).

Remark: *The causality of $H_{fast}(t)$ and $H_{slow}(t)$ present in this model generates a more realistic model than the one proposed by Simoncelli and Heeger (1998), where a Gaussian is used as a temporal profile which is non-causal and inconsistent with V1 physiology.* ■

Figure 5.2 shows the respective spatial and temporal contributions for a V1 simple cell defined as $F^a(\mathbf{x}, t)$.

The spatial parameters of the Gabor function (5.3): θ, f and σ; and the temporal parameter τ of the Gamma function (5.6) define the spatiotemporal orientation of V1 simple cells $F^a(\mathbf{x}, t)$ and $F^b(\mathbf{x}, t)$.

The spatiotemporal orientation of a V1 simple cell is better visualized in the Fourier space (see Section 4.1.3). In the Fourier space the power spectrum of a

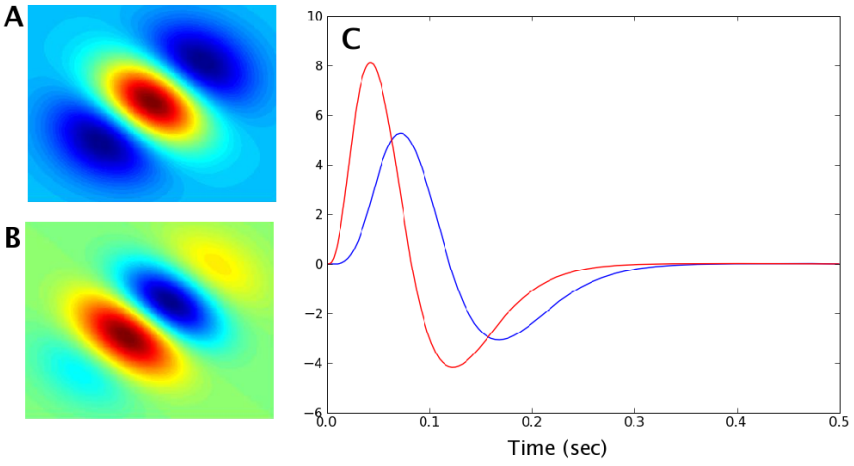


Figure 5.2: Spatial and temporal parts of $F^a(\mathbf{x}, t)$ which models a V1 simple cell. **A-B** are the Gabor derivatives representing the spatial contribution for $F^{even}(\mathbf{x})$ and $F^{odd}(\mathbf{x})$, respectively (see equation (5.4)). **C** is the temporal contribution given by equation (5.7). $H_{fast}(t)$ is represented in red while $H_{slow}(t)$ is represented in blue.

V1 simple cell is described by two blobs centered at $(-\boldsymbol{\xi}_0, \omega_0)$ and $(\boldsymbol{\xi}_0, -\omega_0)$, where $\boldsymbol{\xi}_0 = (\xi_0^x, \xi_0^y)$ and ω_0 are the preferred spatial and temporal frequencies, respectively (see Figure 5.3 (c)). The values of $\boldsymbol{\xi}_0$ and ω_0, which give the spatiotemporal orientation, are given by the input parameters: θ, f, σ and τ (see (5.10)).

The quotient between the highest temporal frequency activation (ω_0) and the highest spatial frequency ($\boldsymbol{\xi}_0$) is the speed of the filter $\mathbf{v} = (v_x, v_y)$, where

$$v_x = \omega_0/\xi_0^x \qquad \text{and} \qquad v_y = \omega_0/\xi_0^y. \tag{5.8}$$

Observing carefully Figure 5.3 (c), it is also possible to see a small activation for the same speed but in the opposite motion direction. The activation in the anti-preferred direction tuning is an effect also seen in real V1-MT cells data (e.g., Snowden et al. (1991)), where V1 cells have a weak suppression in anti-preferred direction (30%) compared with MT cells (92%) (see also Section 3.2.2).

5.1.2 V1 complex cells

As Section 3.1.1 described, some characteristics of V1 complex cells can be explained using a nonlinear combination of V1 simple cells. Their responses are relatively independent of the precise stimulus position inside the receptive field, which suggest a combination of a set of V1 simple cells responses. The complex cells are also invariant to contrast polarity which indicates a kind of rectification of their ON-OFF receptive field responses.

Based on Adelson and Bergen (1985), we define the ith V1 complex cell, located at

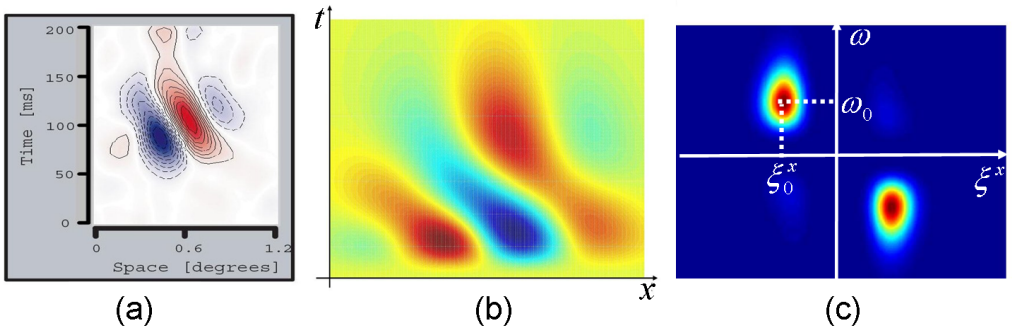


Figure 5.3: (a) Example of a spatiotemporal map of one directionally-selective V1 simple cell (De Valois et al. (2000)). (b) Space-time diagram for $F^a(x,t)$ considering only one spatial dimension x. Here the directionally-selective property is marked and obtained after a linear combination of $F^{even}(x)$, $F^{odd}(x)$, $H_{slow}(t)$ and $H_{fast}(t)$. It is important also to observe the similarities with the biological activation maps measured by De Valois et al. (2000) (a). (c) Spatiotemporal energy spectrum of the directional-selective filter $F^a(x,t)$. The slope formed by the peak of the two blobs (ω_0/ξ_0^x) is the speed tuning of the filter, which will only react for that speed inside a very limited spatiotemporal frequency bandwidth.

$\mathbf{x}_i = (x_i, y_i)$, with spatiotemporal orientation $\boldsymbol{\mu}_i = (\boldsymbol{\xi}_0^i, \omega_0^i)$ as

$$C(\mathbf{x},t) = \big[\big(F^a * L\big)(\mathbf{x}_i,t)\big]^2 + \big[\big(F^b * L\big)(\mathbf{x}_i,t)\big]^2, \tag{5.9}$$

where the symbol $*$ represents the spatiotemporal convolution, and $F^a(\cdot)$ and $F^b(\cdot)$ are the V1 simple cells defined in (5.2). This definition gives independence to stimulus contrast sign and the cell response is constant in time for a drifting sinusoidal as input stimulus. A diagram with the construction of a V1 complex cell defined in (5.9) is shown in Figure 5.4.

Remark: *This complex cell definition does not solve the space invariant property found in real V1 complex cells. By now, this invariance will be taken by MT neurons which have larger receptive fields than V1 neurons.* ■

5.1.3 Spatiotemporal frequency analysis of V1 motion detectors

Thinking about the design of our filter bank, we are interested in the estimation of the spatiotemporal bandwidth of our V1 simple cell model. V1 complex cell $C(\mathbf{x},t)$, defined in (5.9), is completely nonlinear and the Fourier transform cannot be applied in order to analyze its frequency content, that is why we will analyze $F^a(\mathbf{x},t)$ and $F^b(\mathbf{x},t)$. For simplicity and without loss of generality, we will use just one spatial dimension x (in the Fourier domain, we will denote $\boldsymbol{\xi}$ as ξ).

Let us denote by $\tilde{F}^a(\xi,\omega)$ and $\tilde{F}^b(\xi,\omega)$ the respective Fourier transforms of $F^a(x,t)$ and $F^b(x,t)$. Considering an input stimuli $L(x,t) = \delta(x,t)$ the impulse responses are

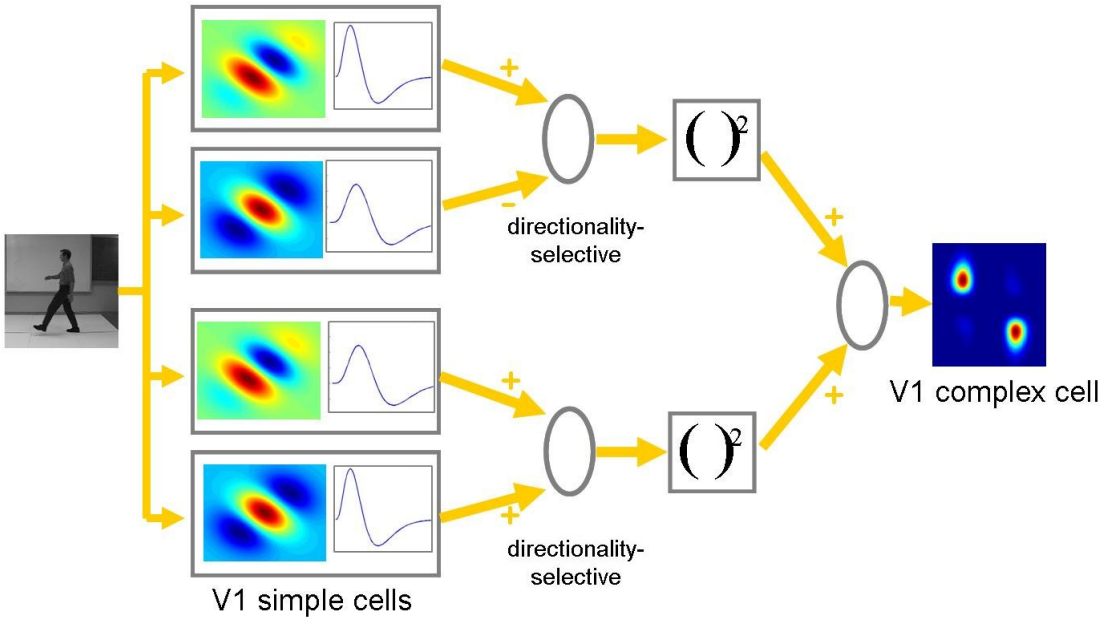


Figure 5.4: V1 complex cell construction following the model described in Adelson and Bergen (1985). The complex cell is created starting from the V1 simple cells defined by (5.2).

defined by

$$\begin{aligned}\tilde{F}^a(\xi,\omega) &= \tilde{F}^{odd}(\xi)\tilde{H}_{fast}(\omega) - \tilde{F}^{even}(\xi)\tilde{H}_{slow}(\omega),\\ \tilde{F}^b(\xi,\omega) &= \tilde{F}^{odd}(\xi)\tilde{H}_{slow}(\omega) + \tilde{F}^{even}(\xi)\tilde{H}_{fast}(\omega),\end{aligned} \tag{5.10}$$

where $\tilde{F}^{odd}$, $\tilde{F}^{even}$, $\tilde{H}_{slow}$ and $\tilde{H}_{fast}$ denote the Fourier transforms of F^{odd}, F^{even}, H_{slow} and H_{fast}, respectively defined by

$$\begin{aligned}\tilde{F}^{odd}(\xi) &= \sigma\sqrt{2\pi}\xi\sinh(\eta\xi\sigma^2)\exp\left(-\frac{\sigma^2(\eta^2+\xi^2)}{2}\right),\\ \tilde{F}^{even}(\xi) &= j\sigma\sqrt{2\pi}\xi^2\sinh(\eta\xi\sigma^2)\exp\left(-\frac{\sigma^2(\eta^2+\xi^2)}{2}\right),\\ \tilde{H}_{fast}(\omega) &= \frac{1}{(1+j\tau\omega)^4} - \frac{1}{(1+j\tau\omega)^6},\\ \tilde{H}_{slow}(\omega) &= \frac{1}{(1+j\tau\omega)^6} - \frac{1}{(1+j\tau\omega)^8},\end{aligned} \tag{5.11}$$

As we previously mentioned, in the Fourier space the power spectrum of a V1 simple cell ($|\tilde{F}^a(\xi,\omega)|^2$ or $|\tilde{F}^b(\xi,\omega)|^2$) is described by two blobs centered at $(-\xi_0^x,\omega_0)$ and $(\xi_0^x,-\omega_0)$, where ξ_0^x and ω_0 are the preferred spatial and temporal frequencies, respectively (ξ_0^x now on denoted as ξ_0). The respective power spectrum $|\tilde{F}^a(\xi,\omega)|^2$ and $|\tilde{F}^b(\xi,\omega)|^2$ are shown in Figure 5.5.

In order to estimate the maximal points, ξ_0 and ω_0, of the power spectrum shown in Figure 5.5, it will be necessary to analyze its derivatives with respect to ξ and ω. Since the analytic solutions of $\partial|\tilde{F}^a(\xi,\omega)|^2/\partial\xi = 0$, $\partial|\tilde{F}^b(\xi,\omega)|^2/\partial\xi = 0$ and

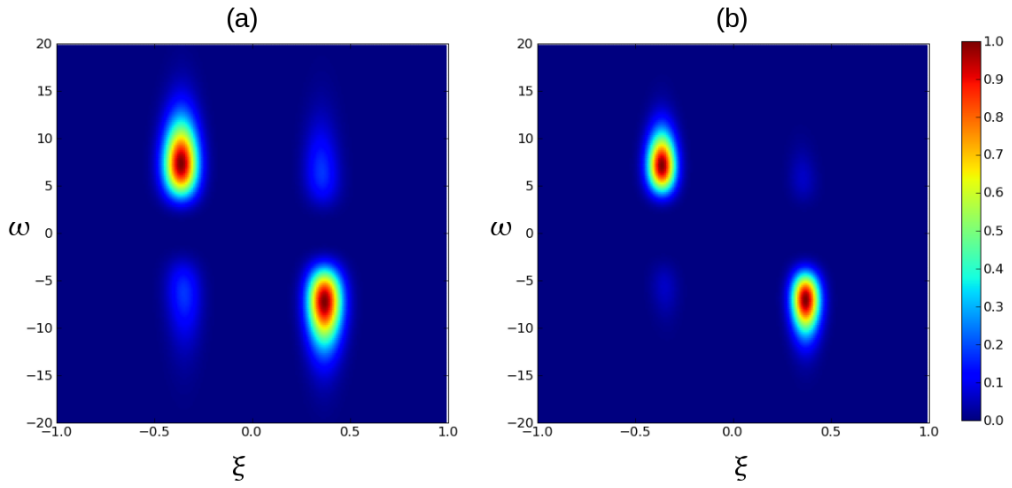


Figure 5.5: (a) Power spectrum of $\tilde{F}^a(\xi,\omega)$. (b) Power spectrum of $\tilde{F}^b(\xi,\omega)$ (b). These graphs were obtained using $f = 0.1$[pixels/cycles], $\sigma = 5.622$ and $\tau = 0.064$[sec]. Numerically we found that $\xi_0 = \pm 0.37$ and $\omega_0 = \pm 7.1$ are the same for both $\tilde{F}^a(\xi,\omega)$ and $\tilde{F}^b(\xi,\omega)$.

$\partial|\tilde{F}^a(\xi,\omega)|/\partial\omega = 0$, $\partial|\tilde{F}^b(\xi,\omega)|/\partial\omega = 0$ do not exist, the values of ξ_0 and ω_0 must be found numerically.

The numerical solution shows that ξ_0 not only depends on the input frequency f, but also on the value of σ defined in (5.3). For a fixed value of τ, the dependency of ξ_0 in $|\tilde{F}^a(\xi,\omega)|^2$ with respect to f and σ is illustrated in Figure 5.6.

The value of σ also defines the orientation selectivity of V1 neurons. Small values of σ originate broad orientation selectivity, while large values of σ improve the orientation selectivity of V1 neurons (see Figure 5.7 **A**-**B**). The orientation selectivity factor (OF) is then defined as the standard deviation of the Gaussian that better fits the orientation selectivity curve. So, defining σ as $\sigma = \sigma_{factor} f$, the relationship between OF and σ_{factor} is illustrated in Figure 5.7 **C**.

Watson and Ahumada (1983) and Watson and Ahumada (1985) proposed a relationship between f and σ as follows: *the diameter of the Gaussian defined in (5.3) at half height must be 1.324 times the period of the function*. This definition gives the cell a spatial frequency bandwidth (at half height) of one octave. In this case, using this relationship the value of σ is fixed as

$$\sigma = \frac{0.5622}{f}, \tag{5.12}$$

where f is the spatial frequency of the Gabor functions defined in (5.3). With this value of σ the curves for ξ_0 and ω_0 depending on f and τ are shown in Figure 5.8. From Figure 5.8 it is possible to conclude that:

- The power spectrum of $\tilde{F}^a(\xi,\omega)$ and $\tilde{F}^b(\xi,\omega)$ have the same ξ_0 and ω_0 values.
- The value of ξ_0 only depends on f.

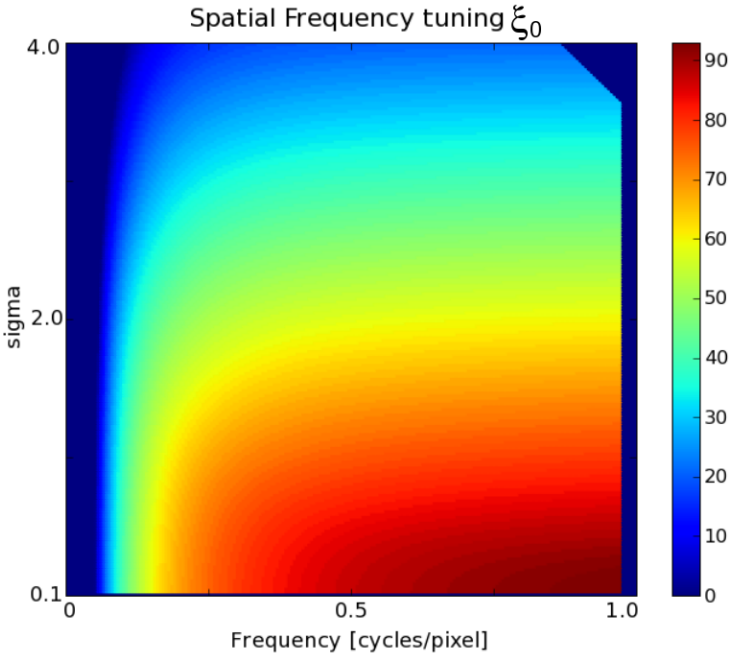


Figure 5.6: Value of the spatial frequency tuning ξ_0 depending on the input parameters f and σ defined in (5.3) (τ fixed). For a fixed value of f, the value of the spatial frequency tuning ξ_0 decreases while σ increases.

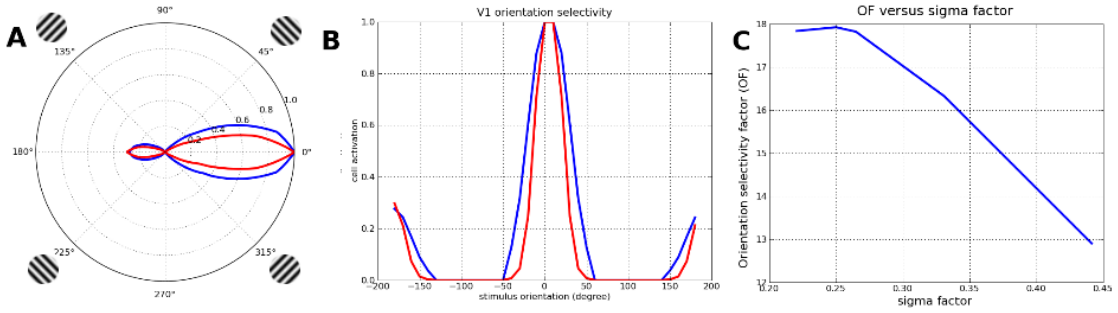


Figure 5.7: σ dependency in the orientation selectivity of a V1 complex cell modeled as (5.9). The orientation selectivity graph was obtained applying different drifting gratings, with different drifting directions, as input stimulus. **A-B** shows the effect of σ in the orientation selectivity of a V1 simple cell for two different values of σ: red curve was obtained for a σ two times larger than the σ used to obtain the blue curve. **C** shows how the orientation selectivity, represented by OF, varies according to the value of σ_{factor}.

- The value of ω_0 only depends on τ.

Specific values of ξ_0 and ω_0 for certain values of f and τ are shown in Table 5.1, different values could be approximately found using linear interpolation.

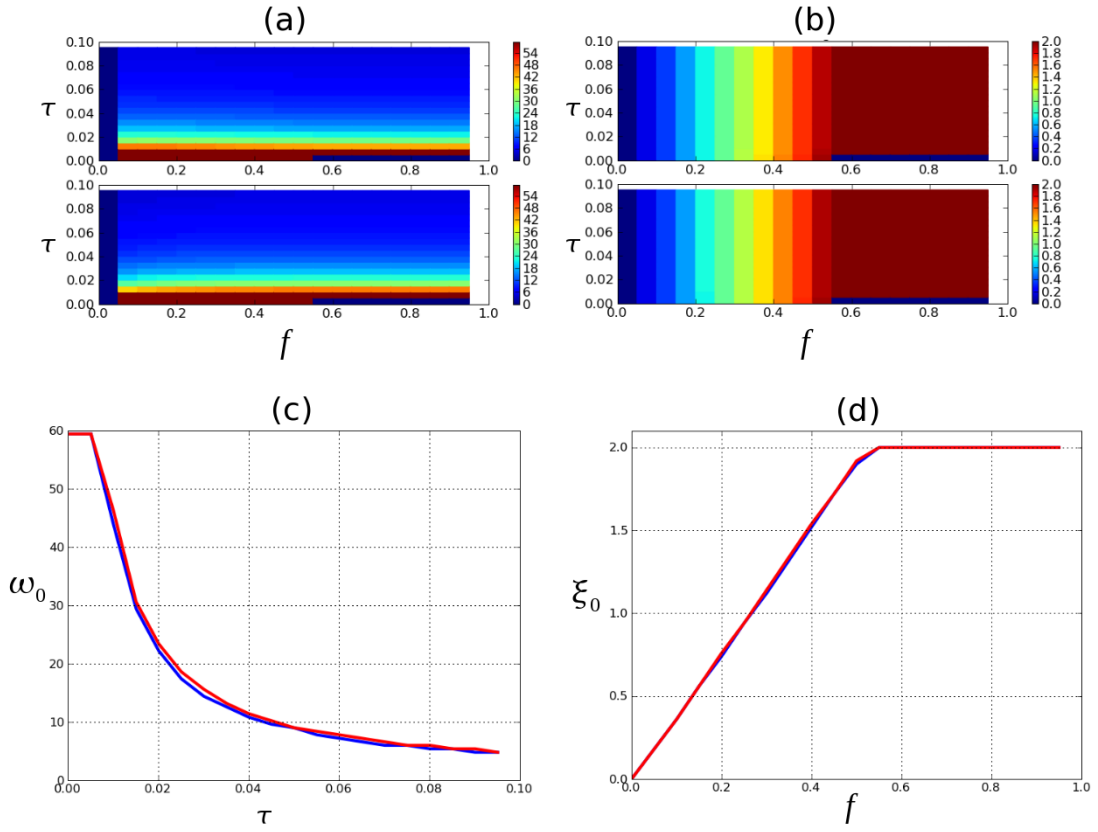


Figure 5.8: Values of ω_0 and ξ_0 as a function of the input parameters f and τ. (a) Surface obtained for ω_0 depending on the parameters f and τ for $|\tilde{F}^a(\xi,\omega)|^2$ (upper graph) and $|\tilde{F}^b(\xi,\omega)|^2$ (lower graph). The surface shows that the value of ω_0 is independent of f. (b) Surface obtained for ξ_0 depending on the parameters f and τ for $|\tilde{F}^a(\xi,\omega)|^2$ (upper graph) and $|\tilde{F}^b(\xi,\omega)|^2$ (lower graph). The surface shows that the value of ξ_0 is independent of τ. (c) Relationship between ω_0 and τ for $|\tilde{F}^a(\xi,\omega)|^2$ (blue) and $|\tilde{F}^b(\xi,\omega)|^2$ (red). (d) Relationship between ξ_0 and f for $|\tilde{F}^a(\xi,\omega)|^2$ (blue) and $|\tilde{F}^b(\xi,\omega)|^2$ (red). Specific values of ω_0 and ξ_0 are displayed in Table 5.1.

To conclude this section, let us remind that the spatiotemporal frequency tuning of a V1 simple/complex cell defines its speed selectivity inside that spatiotemporal bandwidth. These cells are not velocity tuned in the sense defined in Section 3.1.1 where a velocity tuned neuron has a response which is independent of the spatiotemporal frequency content of the input stimulus. In our case, velocity tuned neurons can be constructed adding multiple V1 complex cells sharing the same speed tuning, i.e., with the same quotient ω_0/ξ_0 as it is shown in Figure 5.9.

Table 5.1: Values of ω_0 depending on τ and values of ξ_0 depending on f. Different values can be found using linear interpolation. The relationship between these parameters is plotted in Figure 5.8 (c) and (d)

$\tau[sec]$	$\omega_0[rad/sec]$	$f[cycles/pixel]$	$\xi_0[rad/pixel]$
0.0001	59.4	0.0	0.0
0.0051	59.4	0.05	0.18
0.0101	43.8	0.1	0.36
0.0151	29.4	0.15	0.56
0.0201	22.2	0.2	0.74
0.0251	17.4	0.25	0.94
0.0301	14.4	0.3	1.12
0.0351	12.6	0.35	1.32
0.0401	10.8	0.4	1.52
0.0451	9.6	0.45	1.72
0.0501	9.0	0.5	1.9
0.0551	7.8	0.55	2.0
0.0601	7.2	0.6	2.0
0.0651	6.6	0.65	2.0
0.0701	6.0	0.7	2.0
0.0751	6.0	0.75	2.0
0.0801	5.4	0.8	2.0
0.0851	5.4	0.85	2.0
0.0901	4.8	0.9	2.0
0.0951	4.8	0.95	2.0

5.2 MT BASIC ENTITY

5.2.1 General definition

MT neurons pool incoming information from V1 according to the shape and characteristic of their receptive field. Every V1 neuron lying inside the MT receptive field contributes to the MT cell activation.

In our model, we chose a feedforward connectivity pattern as it is shown in Figure 5.10. Each connected V1 neuron has a respective connection weight. The connection weights are given by the desired tuning values of the MT receptive field.

Let us define the absolute difference of motion direction-selectivity φ_{ij} between the ith and jth neurons as

$$\varphi_{ij} = |\theta_i - \theta_j|,$$

where θ_i and θ_j are the motion direction-selectivity tunings of the ith and jth neurons, respectively.

So, for a MT neuron i, the criteria is to consider all the jth V1 cells inside the MT receptive field, such that $\varphi_{ij} < \pi/4$ radians. The weight associated to the con-

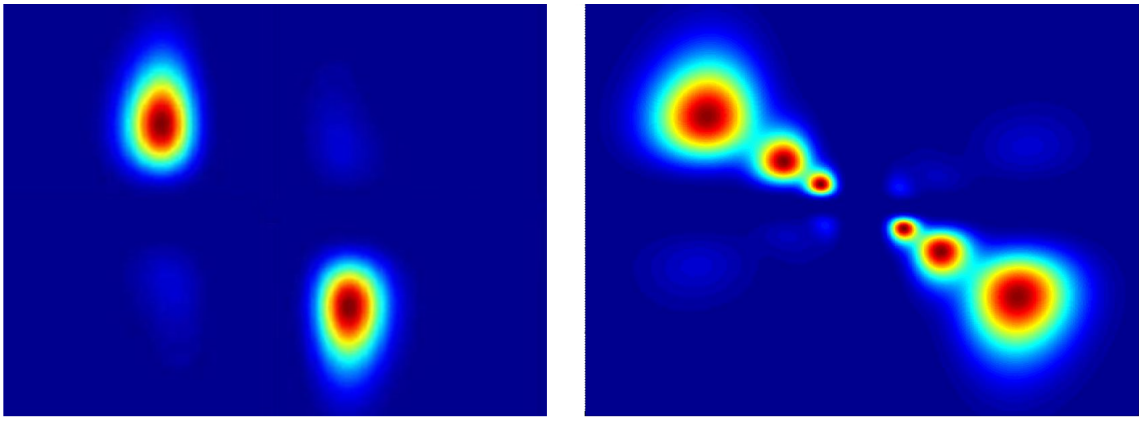


Figure 5.9: V1 velocity tuned neuron built starting from the V1 complex neurons defined in (5.9). (a) Power spectrum of the V1 complex cell defined in (5.9). This neuron has a speed tuning but it is not a velocity tuned neuron. A V1 velocity tuned neuron (b) is defined by adding different V1 complex cells sharing the same speed tuning, i.e., sharing the quotient ω_0/ξ_0.

nection between the V1 pre-synaptic neuron j and the MT post-synaptic neuron i is proportional to the angle φ_{ij} between the two preferred motion direction-selectivity (see Figure 5.11). The connection weight w_{ij} between the jth V1 cell and the ith MT cell is given by

$$w_{ij} = \begin{cases} k_c w_{cs}(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) & \text{if } 0 \leq \varphi_{ij} \leq \frac{\pi}{4}, \\ -k_c w_{cs}(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) & \text{if } \varphi_{ij} > \frac{3\pi}{4}, \end{cases} \tag{5.13}$$

where k_c is an amplification factor, α_{ij} is the absolute angle between the preferred ith MT cell direction and the preferred jth V1 cell direction. $w_{cs}(\cdot)$ is the weight associated to the difference between the center of MT cell $\mathbf{x}_i = (x_i, y_i)$ and the V1 cell center position $\mathbf{x}_j = (x_j, y_j)$. The value of $w_{cs}(\cdot)$ depends on the shape of the receptive field associated to the MT cell.

Negative weights in equation (5.13) (when $\varphi_{ij} > 3\pi/4$) are included to improve the direction selectivity of MT neurons eliminating the two-blob shape obtained for V1 neurons (see Figure 5.7 A). The result of this pooling mechanism (without any other interaction or dynamic) improves the direction selectivity of MT neurons, obtaining only one blob as it is shown Figure 5.12.

In a general frame, the dynamic activation of a MT neuron mainly depends on three variables: the previous activation of the MT neuron, the activation of V1 neurons inside its receptive field Ω and the activation of V1 neurons inside the surround Φ of MT cell. Defining the activation of a MT neuron as $A^{MT}(t)$ and the activation of the jth V1 neuron as $A_j^{V1}(t)$, we propose that $A^{MT}(t)$ is a function (f) of

$$A^{MT}(t) = \mathbf{f}\left(A^{MT}(t-\delta t), \sum_{\Omega} A_j^{V1}(t-\delta t), \sum_{\Phi} A_j^{V1}(t-\delta_\Phi)\right), \tag{5.14}$$

where δt is the time discretization unit used to compute the evolution in time of the

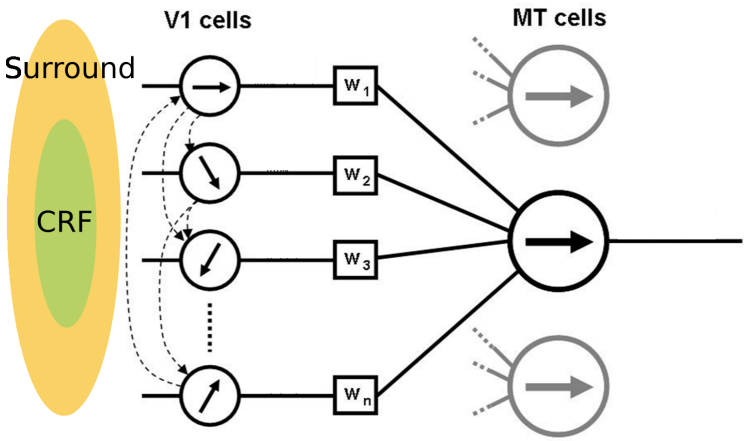


Figure 5.10: V1 neurons connecting to a MT neuron. Each MT cell receives as input the afferent V1 cells. The V1 neurons considered are those lying inside the receptive field of the MT cell. The MT receptive field also defines the connection weights of each V1 neuron.

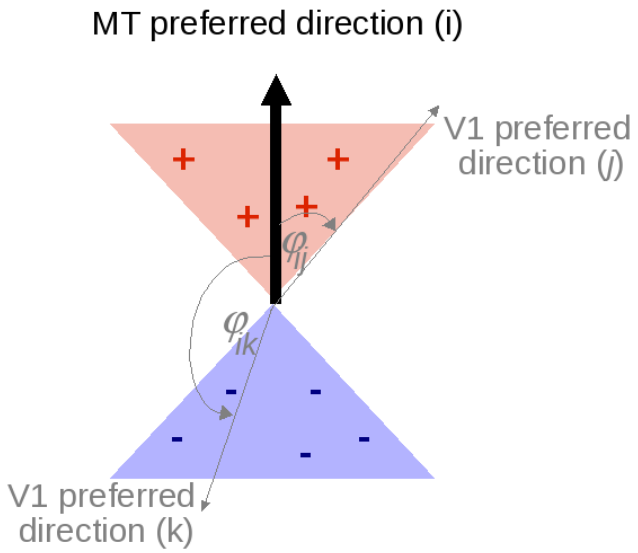


Figure 5.11: The connection weights between V1 and MT neurons are modulated by the cosine of the angle φ_{ij} (φ_{ik}) between the preferred direction of the ith MT neuron and the preferred direction of the jth (kth) V1 neuron. If φ falls into the red zone, the connection weight associated is positive. By the contrary, if φ falls into the blue zone, the connection weight is negative (see equation (5.13)).

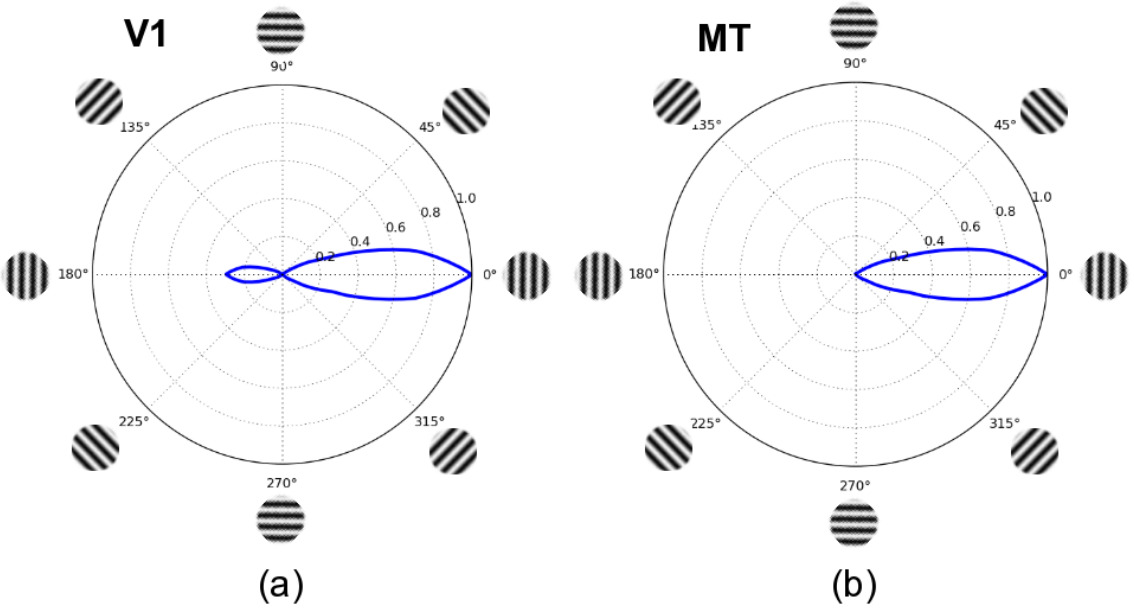


Figure 5.12: (a) Motion direction selectivity of a V1 neuron. (b) Motion direction selectivity of a MT neuron. The orientation selectivity graphs were obtained applying different drifting gratings, with different drifting directions, as input stimulus. The polar diagram for a MT neuron was obtained pooling the responses of V1 neurons with the respective connection weights defined in equation (5.13). The direction selectivity of MT neurons is highly improved compared with the V1 direction selectivity, passing from bimodal to unimodal selectivity.

activity of a MT neuron. δ_Φ is the time delay to consider the activation of V1 neurons belonging to MT surround.

See later: *In Chapters 7, 8 and 9, the dynamics and evolution in time of MT neurons will be defined by their models. The activation of a MT neuron A^{MT} will be, for instance, interpreted as: the value of its membrane potential or the mean firing rate.* ■

5.2.2 MT center-surround interactions

The activation of the MT surround modulates the activation of its classical receptive field. This modulation, given by the third argument of function f in equation (5.14), depends on the shape and organization of the center-surround interactions (Xiao et al. (1997b)) which is usually ignored in MT-like models. In most cases this modulation is inhibitory, but Huang et al. (2007) showed that this interaction, depending on the input stimulus, can be also integrative. The direction tuning of the surround compared with the center tends to be either the same or opposite, but rarely orthogonal (see Section 3.2.3). Considering this, in this thesis we propose different center-surround interactions and different surround geometries.

Following the results found by Born (2000), we consider three types of MT center-surround interactions in our model. Our claim is that the antagonistic surrounds contain key information about the motion characterization, which could highly help in a real application where motion features must be analyzed, such as human action recognition (see Chapters 7 and 8). More precisely, we propose a cell with only the activation of its classical receptive field (CRF) and two cells with inhibitory surrounds as shown in Figure 5.13.

Regarding different surround geometries, we included four types of MT cells (see Figure 5.14): one basic type of cell activated only by its CRF, and three other types with inhibitory surrounds. We claim that the information obtained thanks to the asymmetric surrounds brings complementary information about motion (such as, motion contrasts), and we will illustrate this in the Part II of this thesis. The tuning direction of the surround is always the same as the CRFs, but their spatial geometry changes, from symmetric to asymmetric-unilateral and asymmetric-bilateral surround interactions. Of course, it is important to mention that this approach remains a coarse approximation of the real receptive field shapes.

5.3 Implementation of V1-MT as network of neurons

For all the approaches presented in this thesis, V1 and MT neurons are arranged as V1 and MT layers of neurons, respectively. Those layers are able to process the

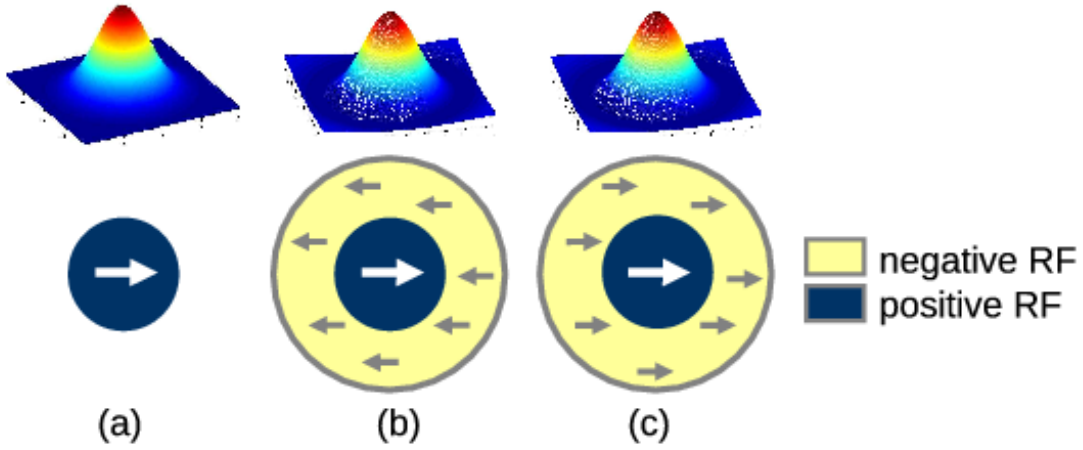


Figure 5.13: Center-surround interactions modeled in the MT cells. (a) Classical receptive field (CRF) modeled through a Gaussian. (b)-(c): Two receptive fields with inhibitory surround, which are modeled with a Difference-of-Gaussians (DoG). (b) Inhibitory surround with antagonistic direction tuning compared to the CRF. (c) Inhibitory surround with the same direction tuning than the CRF.

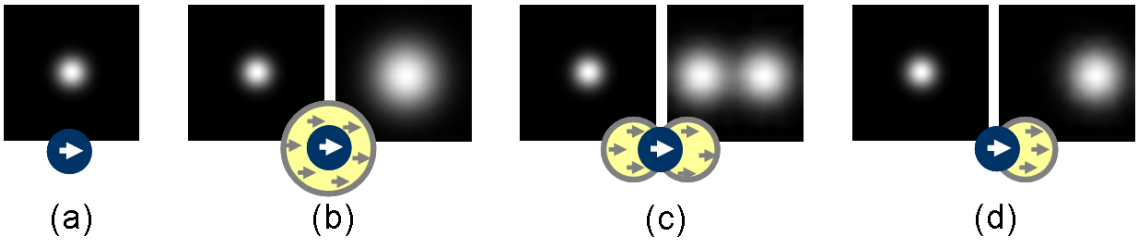


Figure 5.14: MT center-surround geometries modeled in our approach. (a) Classical receptive field CRF modeled with a Gaussian. All the surrounds from (b) to (d) are also modeled by Gaussians. (b) Represents a symmetric surround. The two groups of cells with asymmetric surrounds are represented in (c) and (d). (c) Represents a bilateral asymmetric surround. (d) Represents a unilateral asymmetric surround. There is an important presence of anisotropic surround interactions in MT cells: In Xiao et al. (1997b, 1995), the authors showed that within the MT cells with surround suppression, the configuration (b) is present only in the 25% of the cells, while (c) and (d) cover the resting percentage with a presence of 50% and 25%, respectively.

motion information contained inside a defined visual field. This section describes how V1 and MT layers are defined.

5.3.1 Organization of V1 layers

Given V1 complex cells modeled by (5.9), we consider N_L layers of V1 cells (see Figure 5.15). Each layer is built with cells sharing the same speed tuning $\mathbf{v} = \omega_0/\|\boldsymbol{\xi}_0\|$ and N_{or} different spatial orientations $\theta_i = \arctan(\xi_0^{yi}/\xi_0^{xi}), \quad i = 1, .., N_{or}$. All the V1 cells belonging to one layer, with receptive fields centered in the position (x_i, y_i), form what we call a *column*. One *column* has as many elements as the number of orientations defined N_{or}. See Figure 5.15 for an illustration.

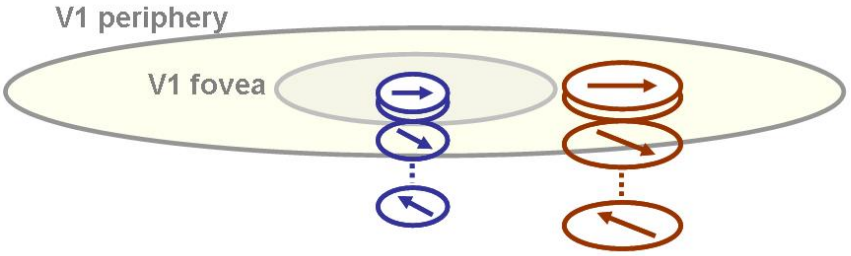


Figure 5.15: Diagram with the architecture of one V1 layer. There are two different regions in V1, the fovea and periphery. Each element of the V1 layer is a column of N_{or} V1 cells, where N_{or} corresponds to the number of orientations.

The centers of the receptive fields are distributed along a radial log-polar scheme with a foveal uniform zone. The related one-dimensional density $d(r)$, depending of the eccentricity r, is defined by

$$d(r) = \begin{cases} d_0 & \text{if } r \leq R_0, \\ d_0 R_0 / r & \text{if } r > R_0. \end{cases} \tag{5.15}$$

The cells with an eccentricity r less than R_0 have an homogeneous density and their receptive fields refer to the retina fovea (*V1 fovea*). The cells with an eccentricity greater than R_0 have a density depending on r and receptive fields lying outside the retina fovea (*V1 periphery*). An schematic representation is shown in Figure 5.17 (a).

Reminder: *V1 complex cells defined in* (5.9) *are not velocity tuned neurons. They are sensible to velocity but inside a very limited spatiotemporal frequency bandwidth defined by the input parameters of* (5.3)*: τ. f and θ (see Figure 5.9 (a)).* ■

Along this thesis we will not be concerned about speed, but only about the direction of motion. So, in our case for a given spatial orientation θ we are interested into pave the spatiotemporal frequency space of interest in an homogeneous manner. Following the work done by Mante and Carandini (2005), our frequency space of interest is limited to: spatial frequency range of 0.05 to 0.2 cycles/pixel, and temporal

frequency range of 2 to 8 Hz. Inside this frequency space we used three different spatial frequencies: 0.05, 0.1 and 0.2 cycles/pixel; and three different temporal frequencies: 2, 4 and 8 Hz. Using these values, for a given spatial orientation θ the spatiotemporal frequency space of sensibility is shown in Figure 5.16.

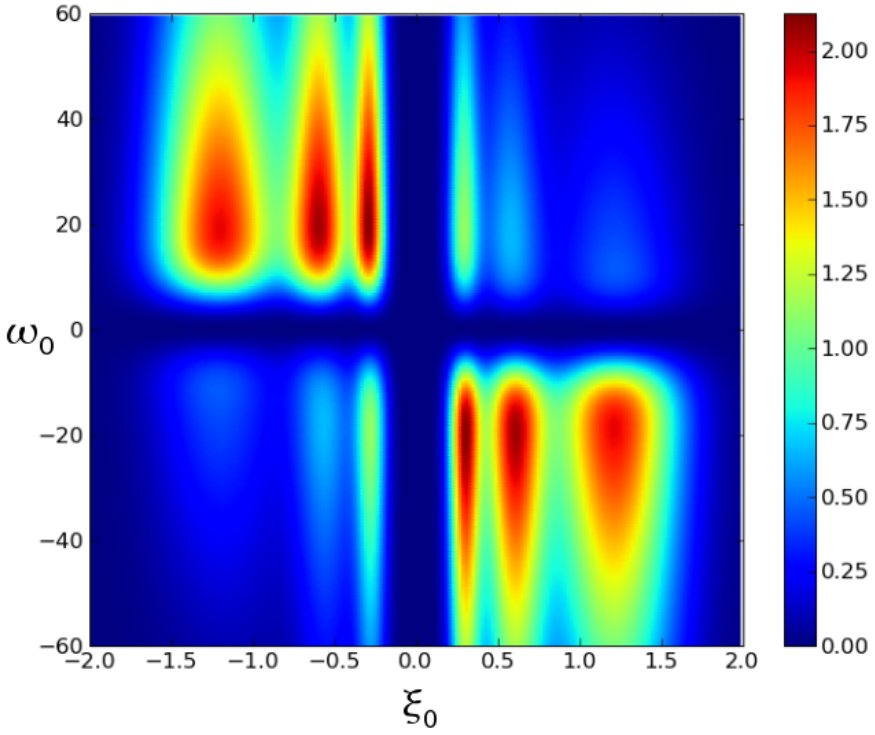


Figure 5.16: Frequency space tiled by the different V1 complex cells used in this thesis. This graph was obtaining combining nine V1 complex cells with spatial and temporal frequency tuning of $\{0.05, 0.1, 0.2\}$[cycles/pixel] and $\{2, 4, 8\}$[Hz], respectively. The values of the parameters f and τ of (5.9) were obtained from Table 5.1.

5.3.2 Organization of MT layers

Analogous to V1 cells, MT cells are distributed in a log-polar architecture, with a homogeneous area of cells in the center and a periphery where the density decreases with the distance to the center of focus. While the density of cells decreases with the eccentricity, the size of the receptive fields increases preserving its original shape. Figure 5.17 (b) shows an example of the log-polar distribution of MT cells. Each center-surround interaction or center-surround geometry defines a MT layer.

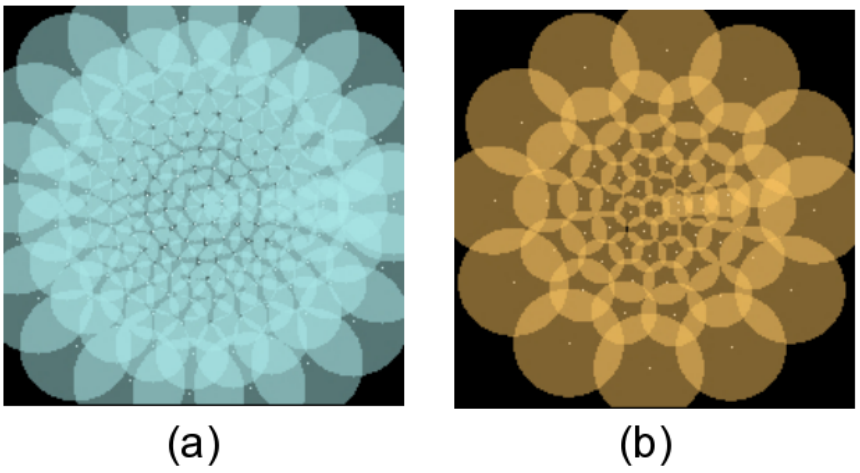


Figure 5.17: Sample of log-polar architecture used for V1 and MT layers, showing the difference of cells density between V1 (a) and MT (b). The cell distribution law is divided into two zones, a homogeneous distribution in the center with a certain radius and then a periphery where the density of cells decays with the eccentricity.

Part II

Human Action Recognition

CHAPTER 6

STATE OF THE ART OF HUMAN ACTION RECOGNITION

"Never confuse motion with action"
–Benjamin Franklin (1706-1790)

Contents

OVERVIEW

Human action recognition can be defined as the process of labelling input video sequences with actions. This is a challenging visual task which has been vastly studied in several communities.

Establishing an automatic human action recognition system is extremely challenging. In general, many constraints and assumptions are needed in order to obtain satisfactory results. Let us mention some difficulties: For example, the same action, performed from different points of view, can lead to very different image observations; Different persons can appear differently due to differences in anthropometry, clothes, skin color; The appearance can also be influenced by the lightning, specially if it is not homogeneous inside the image; Another factors can come from the distance camera-target, the speed of the action performed or the localization inside the video.

In this chapter we briefly describe the state of the art of the human action recognition problem. We also mention some notions about how some parts of the brain could be involved in this visual processing task.

Keywords: Human action recognition, biological motion recognition, point-light stimuli, fMRI.

Organization of this chapter

Section 6.1 describes the state of the art of human action recognition in the computer vision community. Section 6.2 describes studies done in neuroscience trying to understand the underneath mechanism of the human action recognition task. Finally, Section 6.3 describes about the contributions done in the computational neuroscience community.

6.1 How computer vision does it?

Human action recognition and human motion analysis in real scenes remain challenging problems in computer vision and it has been vastly studied in the last 20 years (Aggarwal and Cai (1999); Gavrila (1999); Moeslund et al. (2006); Poppe (2007)). Human action recognition is closely related to different research lines of human motion analysis that will no be treated in this chapter, such as, gesture recognition or hand-pose estimation (see Mitra and Acharya (2007) and Erol et al. (2007), respectively).

Motion is the key feature for a wide class of computer vision approaches. Existing methods consider different motion representations or characteristics, such as coarse motion estimation, global motion distribution, local motion feature detection or spatiotemporal structure learning (Zelnik-Manor and Irani (2001); Efros et al. (2003); Laptev et al. (2007); Dollar et al. (2005); Niebles et al. (2008); Wong et al. (2007)).

Human motion can be interpreted of different manners, for instance, the hierarchical methodology proposed by Moeslund et al. (2006) describes a human action as: action primitive, action and activity. Action primitive is an atomic movement as for example "raise up left arm". A collection of action primitives can describe an action, such as, "walking". Finally, activities is a series of actions, which give the interpretation of the action performed.

In general, the human motion can be interpreted into two categories (see Figures 6.1 and 6.2):

1. **Holistic representations**, where the visual information is encoded as a whole. Within holistic representations, we can account: the analysis of the shape of the silhouette evolution across time (Bobick and Davis (2001); Blank et al. (2005); Wang and Suter (2007); Mokhber et al. (2008)); the analysis of the optical flow inside the ROI (Efros et al. (2003)); grid-based representation as a combination of local descriptors (Ragheb and Hancock (2003); Zhu et al. (2006); Tran and Sorokin (2008); Thurau and Hlavac (2008)); 3D spatiotemporal volumes (Blank et al. (2005); Ogata et al. (2006); Gorelick et al. (2007); Yilmaz and Shah (2008); Jiang and Martin (2008)); generic human model recovery (Hogg (1983); Rohr (1994); Goncalves et al. (1995)).

2. **Path-based representations**, where the visual information is encoded as a collection of small, independent patches. The lack of a background model, proper localization of targets and partial occlusion interfers with the estimation of the ROI. In these cases, the path-based approaches present a valable alternative. The patches (2D or 3D) can be treated independently and each action class is described by a distribution over all the patches. Within path-based approaches we can cite: space-time interest points (Dollar et al. (2005); Laptev et al. (2007));

extraction of motion periodicity characteristics (Polana and Nelson (1997); Seitz and Dyer (1997); Cutler and Davis (2000); Collins et al. (2002)); grouping patches into a codebook representation (Chomat et al. (2000); Lowe (2004); Jhuang et al. (2007); Escobar and Kornprobst (2008); Schindler and Van Gool (2008)); grid-based representation (Laptev and Perez (2007); Laptev et al. (2008); Fathi and Mori (2008)); correlations between features (Fanti et al. (2005); Wong et al. (2006); Wong and Cipolla (2007); Niebles and Fei-Fei (2007); Kim et al. (2007); Liu et al. (2008); Liu and Shah (2008)); body parts modeling and tracking (Gavrila and Davis (1996); Shah and Jain (1997); Gavrila (1999)).

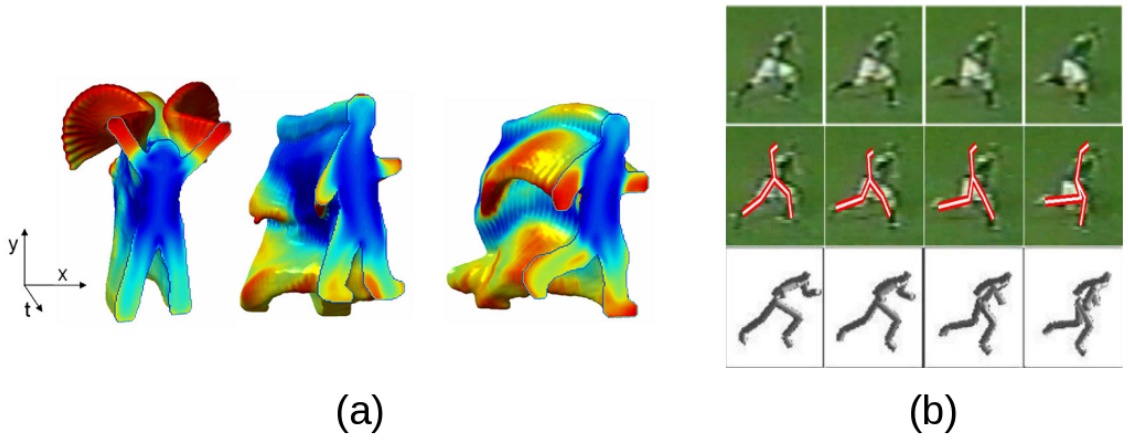


Figure 6.1: Samples of holistic representations for human motion analysis. (a) Examples of the local shape-time saliency features of Blank et al. (2005). (b) Skeleton extraction from the input sequence of the top row of Efros et al. (2003).

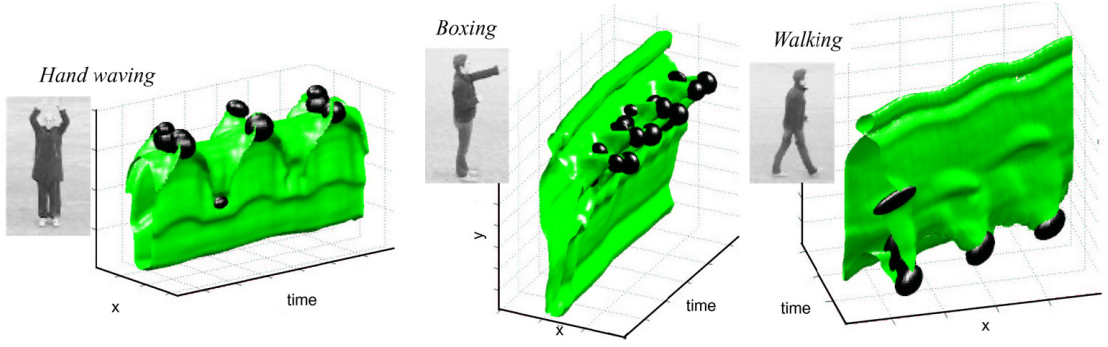


Figure 6.2: Sample of a path-based representation proposed by Laptev et al. (2007). The figure shows examples of scale and velocity adapted local motion events for three different actions: *hand-waving, boxing* and *walking*. Events are illustrated as dark ellipsoids and correspond to corners in a 2D+t representation of the shape moving.

An important category of approaches mentioned in the previous paragraph is based on the motion information. For example, it was shown that a rough description of motion (in Efros et al. (2003)) or the global motion distribution

(Zelnik-Manor and Irani (2001)) can be successfully used to recognize actions. Local motion cues are also widely used. For example, in Laptev et al. (2007), the authors propose to use event-based local motion representations (here, spatial-temporal chunks of a video corresponding to 2D+t edges) and template matching. The idea of extracting spatiotemporal features has been proposed in several contributions such as Dollar et al. (2005), and then Niebles et al. (2006); Wong et al. (2007), using the notion of cuboids. Another stream of approaches was inspired by the work of Serre (2006), first applied to object recognition (Serre et al. (2005); Mutch and Lowe (2006)) and then extended to action recognition (Sigala et al. (2005); Jhuang et al. (2007)).

6.2 How the brain does it?

Action recognition has been addressed in psychophysics where remarkable advances were made in the understanding of human action perception (Blake and Shiffrar (2007)). The perception of human action is a complex task that combines not only the visual information, but additional aspects such as social interactions or motor system contributions. From several studies in psychophysics, it has been shown that our ability to recognize human actions does not need necessarily a real moving scene as input. In fact, we are also able to recognize actions when we watch some point-light stimuli corresponding to joint positions (see Figure 6.3, also Johansson (1973)). This kind of simplified stimuli, known as *biological motion*, was highly used in the psychophysics community in order to obtain a better understanding of the underlying mechanism involved.

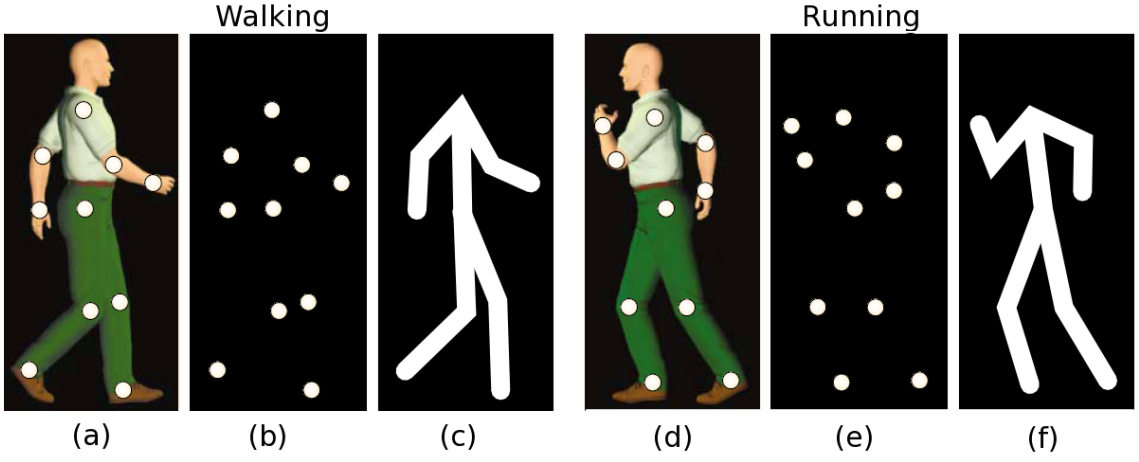


Figure 6.3: Snapshots of two different actions: *walking* (a-c) and *running* (d-f), the junctions to extract point-light stimulus are marked on the figures (a) and (c). For walking: (b) point-light stimulus obtained from (a); (c) stick-figure stimulus obtained from (a). Analogous for running sequence (image adapted from Giese and Poggio (2003)).

The neural mechanisms, processing *form* or *motion* taking part of biological motion recognition, remain unclear. On the one hand, Beintema and Lappe (2002) sug-

gested that biological motion can be derived from dynamic *form* information of body postures and without local image motion. On the other hand, Casile and Giese (2003) proposed a new type of point-light stimulus showing, that only the *motion* information is enough and the detection of specific spatial arrangements of *opponent-motion features* can explain our ability to recognize actions. Finally, Casile and Giese (2005) showed that biological motion recognition can be done with a coarse spatial location of the mid-level optic flow features.

Interestingly, it was confirmed that in the visual system the motion pathway is also very much involved in the action recognition task (see, e.g., Pucel and Perret (2003); Hirai and Hiraki (2006)), but of course other brain areas (e.g., the form pathway) and mechanisms (e.g., top-down attentional mechanisms) are also involved to analyze complex general scenes.

This dichotomy between motion and form finds some neural basis in the brain architecture and it was confirmed by fMRI studies such as Grossman et al. (2000); Vaina et al. (2001); Michels et al. (2005). A simplified representation of the visual processing is that there exists two distinct pathways: the *dorsal* stream (motion pathway) with areas such as V1, MT, MST, and the *ventral* stream (form pathway) with areas such as V1, V2, V4. Both of them seem to be involved in the biological motion analysis.

6.3 Existing bio-inspired models

Nowadays, we can observe a special interest for the so-called bio-inspired approaches to model a part of the visual system functionalities. For example, we presented in Section 4.2 several bio-inspired models for motion estimation. The bio-inspiration term comes from the modelization of a system following the hierarchical architecture of the visual system, and not only its architecture, but also different functionalities. In the context of human action recognition, which implies much more complex processing, some degree of simplification and abstraction is needed.

Because the problem by itself is also very challenging, there are, up to our knowledge, very few bio-inspired approaches. In this section, we present the two main bio-inspired models existing in the literature.

6.3.1 Giese and Poggio's model

Giese and Poggio (2003) proposed a model for visual processing in the dorsal (*motion*) and ventral (*form*) pathways. The stages of their model performing motion processing are described in Section 4.2.2, and a diagram of their model was shown in Figure 4.12. They propose a motion pattern neuron created from snapshot neurons' outputs. The snapshot neurons involved in the encoding of the same motion pattern are summed in a motion pattern neuron (see Figure 6.4). Snapshot neurons have asymmetric lat-

eral connections that pre-excites temporally subsequent snapshot neurons encoding the same body configuration. Temporally, previous snapshot neurons are inhibited. The motion pattern neurons have a significantly activity only when the individual snapshot neurons are activated in the correct temporal order.

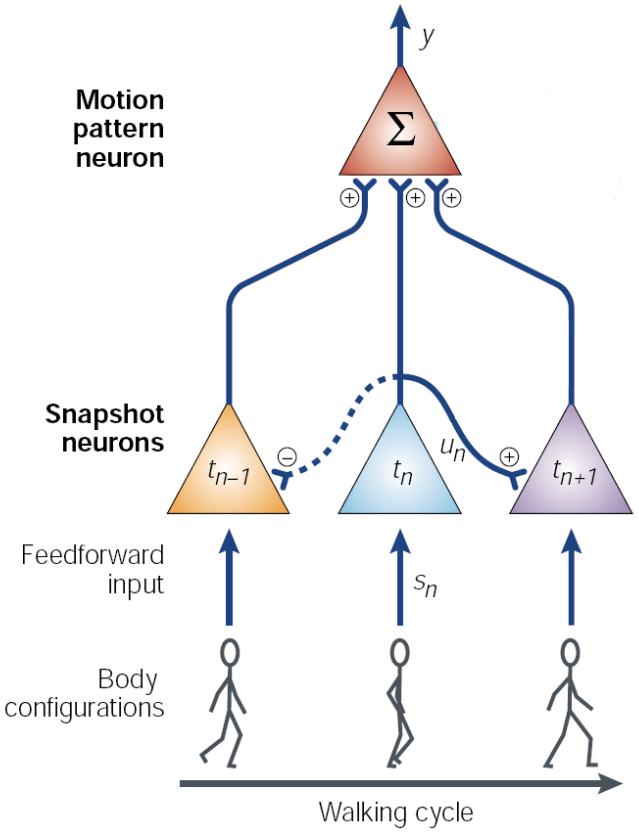


Figure 6.4: Motion pattern neuron created from the snapshot neurons' output encoding the same motion pattern, as e.g., walking. Snapshot neuron at the center (blue) has lateral inhibitory connections with the temporally previous snapshot neuron and excitatory connections with the temporally subsequent snapshot neuron (Image taken from Giese and Poggio (2003)).

Snapshot neurons are ruled by the following differential equation

$$\tau_u \frac{du(t)}{dt} + u_n(t) = s_n(t) + \sum_m w(n-m) f(u_m(t)), \tag{6.1}$$

where u_n is the membrane potential of the nth snapshot neuron; s_n is the output of the radial basis functions trained with learned snapshots; w is the asymmetric lateral coupling strength; f is a sigmoidal nonlinear function and τ_u a time constant.

Similarly, motion pattern neurons are ruled by

$$\tau_y \frac{dy(t)}{dt} + y(t) = \sum_n f(u_n(t)), \tag{6.2}$$

where y is the output of the motion pattern neuron and τ_y its respective time constant.

Motion pattern neurons for the dorsal pathway are created using the same mechanism described for the ventral pathway. The difference is that now the optic-flow pattern neurons are used instead of snapshot neurons (see Figure 4.12).

The motion pattern neurons coming from motion and form pathways were independently tested for the biological motion recognition task, i.e., using point-light stimuli and stick figures, both obtained from real sequences. For stick figures they found that motion pattern neurons coming from form and motion pathways correctly respond to the right sequence, but motion pattern neurons from motion pathway also have a non neglected activation for distractor patterns. This effect was not reproduced when stick figures were replaced by point-light stimuli to test robustness. The motion pathway of the model generalizes from full-body (stick figure) to point-light stimuli because the optic-flow field induced by point-light stimuli is a kind of sampled version of the optic-flow field generated by the stick figure, obtaining of this way a successful recognition. No such generalization occurs in the form pathway, where those motion pattern neurons were not able to recognize the action performed by the point-light stimuli.

They also degraded the quality of point-light stimuli, i.e., they eliminated some junctions. Removing elbows and the feet is specially damaging for recognition, suggesting the crucial role of the opponent motion units present in the motion pathway model.

An extension of this model was proposed by Sigala et al. (2005). This extension only uses the information of the dorsal stream, proposing a biological motion recognition system using a neurally plausible memory-trace learning rule.

The model presented by Giese and Poggio (2003) exhibits several interesting properties for biological motion pattern recognition, such as, spatial and temporal scale invariance, robustness to noise added to point-light motion stimuli and so on. Within its simplifications we can account: no attentional mechanisms, no interaction between dorsal and ventral pathway and no biological inspiration to extract the optical-flow in the first stage of the model. In practice, if we want to consider a new action, new parameter fitting is required to train the respective snapshot and optic-flow neurons.

6.3.2 Jhuang et al.'s model

More recently, Jhuang et al. (2007) proposed a feedforward architecture, which can be seen as an extension of Serre et al. (2005) to treat the action recognition problem in real sequences. They proposed a hierarchical structure based on Giese and Poggio (2003); Serre et al. (2005) and Mutch and Lowe (2006) to obtain as output a feature vector representing the input grayscale video sequences. The diagram of their model is shown in Figure 6.5.

The first stage of the model S_1 is the extraction of motion features. Motion features

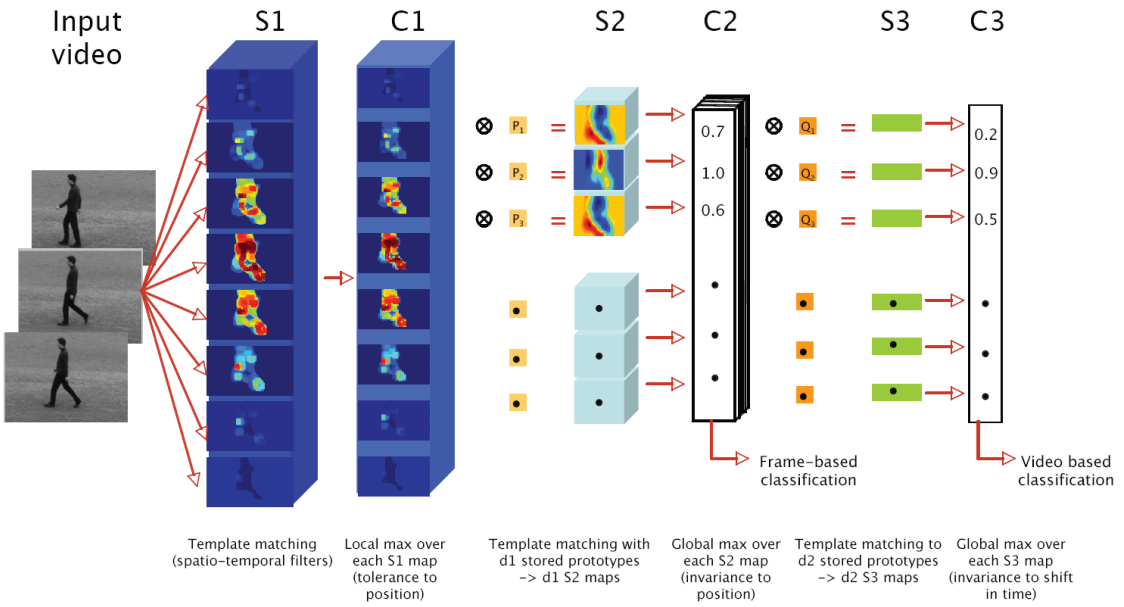


Figure 6.5: Architecture of the model proposed by Jhuang et al. (2007) to perform action recognition in real sequences, see text for details (Image taken from Jhuang et al. (2007)).

are obtained by motion-sensitive units claiming similarity with V1 and MT neurons[1]. Three types of motion-sensitive neurons were tested:

1. **Space-time-gradient based S_1 units**. The motion information is extracted computing the ratio of the temporal gradient to a spatial gradient. They considered the absolute values of gradients in order to have contrast reversal independence.

2. **Optical-flow-based S_1 units**. Direction tuning curve of V1 neurons were modelated as a circular-Gaussian-like function. V1 neurons were grouped into ranges of speeds to define MT neurons. Finally, S_1 units were created combining V1 and MT neurons in a multiplicative way.

3. **Space-time-oriented S_1 units**. The motion information is extracted using a set of spatiotemporal oriented filters. A set of 3D Gabor filters were used to extract the image flow. MT neurons were modeled by 3D Gaussian derivative filters (3rd derivative).

The tolerance to spatial translation is then performed by C_1 units. C_1 units pool the maximum responses from S_1 units over local spatial positions. The resulting C_1 frames are smaller than S_1 frames due to pooling mechanism.

The temporal-prototype-sensitive units S_2 are a prediction of the model and the authors claim that these units are similar to MST neurons. At every position in the

[1]According to their methodology and Chapter 4, only the space-time-oriented S_1 units can be considered as biologically inspired.

C_1 layer, they perform a template matching operation between the current patch C_1 units centered at that position and each of the template d_1 that was previously obtained during training phase. Subsequently, C_2 units perform a maximum-pooling operation adding more position invariance. The maximum extraction is calculated per frame. Stacking all the C_2 responses of a frame, they obtain a vector representation.

The sequence selectivity units S_3 are in charge of to respect the temporal order of the frames for each action. S_3 is then obtained extracting convolving the stored temporal prototype with a C_2 matrix. C_2 matrix is obtained aligning C_2 vectors into columns to create a matrix, Then, at a random column, all the rows inside a temporal window form the C_2 matrix.

C_3 units still add invariance to shifts in time by a maximum-pooling operation. C_3 pools the global maximum across all the pixel position of an input S_3 map, resulting in a scalar representation.

Finally, the classification stage is done with support vector machine (SVM).

Similarly to the work of Giese and Poggio (2003), the model presented by Jhuang et al. (2007) implemented spatial and temporal invariance. The invariance to spatial and temporal scale is achieved considering as many motion detector layers as the number of spatial and temporal scales to be detected, and then applying the $\max$ operator. This model requires pre-processed videos as input, where the action is segmented and the background subtracted. No attentional mechanism or feedback is implemented.

The methodology of the model presented by Jhuang et al. (2007) will be further analyzed and their results compared with the results obtained with the approach proposed in this thesis (see Sections 7.3 and 8.4.3).

CHAPTER 7

ANALOG MODEL IMPLEMENTATION

"You have to see the pattern, understand the order and experience the vision"
–Michael E. Gerber

Contents

OVERVIEW

A wide class of computer vision approaches dealing with human action recognition task are based on motion analysis. Also, fMRI studies have shown that the *dorsal* pathway is very much involved in the action recognition task.

Given the V1-MT feedforward core architecture defined in Chapter 5, we investigated here if such a bio-inspired model can be successfully used to implement a platform performing human action recognition task.

The performance of this architecture is tested using the Weizmann database[1]. We show that modeling different center-surround interactions of MT neurons the recognition performance is significantly improved. We also show a comparison of our results with the results obtained by Jhuang et al. (2007).

Contributions of this chapter:

1. The implementation of an analog V1-MT feedforward architecture to be applied in a real application.
2. The implementation of a classification method to analyze MT analog output.
3. Study of the effect of different MT center-surround interactions in human action recognition performance.

Keywords: Human action recognition, mean motion map, MT center-surround interactions.

Organization of this chapter

This chapter is organized as follows. Section 7.1 describes the specific definitions of V1 and MT neurons. Section 7.2 describes how the *mean motion maps* are defined in order to represent the motion information contained in the input stimulus. It also describes the distance used to compare the similitude between two *mean motion maps*. Finally, Section 7.3 describes the experimental protocol and the results obtained on the Weizmann database.

7.1 ANALOG V1-MT ARCHITECTURE

In Chapter 5 we described our V1-MT feedforward core architecture consisting of a V1 motion detector and a basic MT entity. Now, starting from those definitions, we specify in detail the model for neurons of V1 and MT. In this chapter, neurons work in

[1] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

an analog manner so that the activation of V1 and MT neurons, i.e. the mean firing rate, will be estimated from the value of the membrane potential after a nonlinearity.

7.1.1 V1 neuron implementation

It is well known in biology that V1 neurons' output show several nonlinearities due to: response saturation, response rectification, or contrast gain control (see e.g., Albrecht et al. (2004)). These nonlinearities are reflected in the mean firing rates of those V1 neurons.

Starting from the membrane potential, one of the classical ways to estimate the mean firing rate and thus obtain nonlinear saturation in the V1 response, is to pass the membrane potential through a sigmoid function $S(\cdot)$ defined as

$$S(x) = \left[1 + \exp\left(-a(b - x)\right)\right]^{-1}, \tag{7.1}$$

where the parameters, a and b define the respective slope and horizontal position of the sigmoid function, respectively (see Figure 7.1).

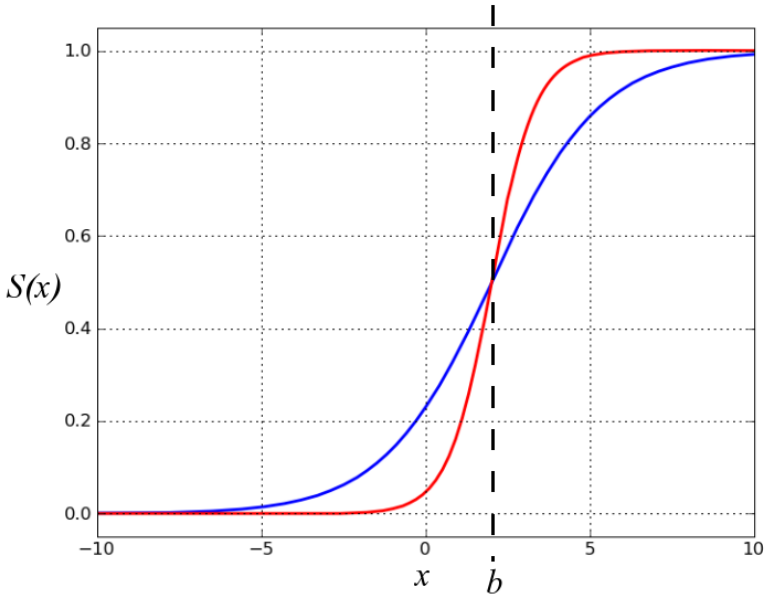


Figure 7.1: Sigmoid function used to convert the membrane potential of the ith V1 neuron (in this case modeled as the output of the V1 complex cell C) to its firing rate $r_i^{V1}(t)$. This nonlinear function serves to model nonlinearities seen in real V1 neurons (see, e.g., Albrecht et al. (2004)). Two graph are shown for different values of the parameters a and b. Red curve: $a = 1.5$, $b = 2$, blue curve: $a = 0.6$, $b = 2$.

In this analog version of our motion processing model, V1 neurons are directly obtained from the output of the complex cell described in Section 5.1.2. We assume that the membrane potential of the ith V1 neuron is modeled by the output of the complex cell $C(\cdot)$ defined in equation (5.9). So, the mean firing rate $r_i^{V1}(t)$ of the ith

V1 neuron, is estimated by

$$r_i^{V1}(t) = S\left(C(\mathbf{x}, t)\right), \tag{7.2}$$

where the parameters a and b of the sigmoid function were tuned to have a suitable response in the case of drifting gratings as inputs.

7.1.2 MT neuron implementation

Let us now precise the definition of a MT neuron given in equation (5.14), where its activity $A^{MT}(t)$ was defined by

$$A^{MT}(t) = \mathbf{f}\left(A^{MT}(t-\delta t), \sum_{\Omega} A_j^{V1}(t-\delta t), \sum_{\Phi} A_j^{V1}(t-\delta_\Phi)\right).$$

Along this chapter, the activity of a MT neuron $A^{MT}(t)$ will be modeled by its membrane potential $u^{MT}(t)$.

In this work, we chose that MT neurons are modeled by a simplified conductance-based neuron without input currents (see e.g., Gerstner and Kistler (2002); Destexhe et al. (2003))[2]. Considering a MT neuron i, its membrane potential $u_i^{MT}(t)$ evolves in time according to the conductance-driven equation:

$$\begin{aligned}\tau\frac{du_i^{MT}(t)}{dt} = {} & G_i^{exc}(t)\left(E^{exc} - u_i^{MT}(t)\right) + G_i^{inh}(t-\delta)\left(E^{inh} - u_i^{MT}(t)\right) \\ & + g^L\left(E^L - u_i^{MT}(t)\right),\end{aligned} \tag{7.3}$$

where E^{exc}, E^{inh} and $E^L = 0$ are constants with typical values of 70mV, -10mV and 0mV, respectively. According to equation (7.3), $u_i^{MT}(t)$ will belong to the interval $\left[E^{inh}, E^{exc}\right]$ and it will be driven by several influences:

- The first term of equation (7.3) refers to input pre-synaptic neurons and it will push the membrane potential $u_i^{MT}(t)$ towards E^{exc}, with a strength defined by $G_i^{exc}(t)$.
- The second term of equation (7.3) also coming from pre-synaptic neurons will drive $u_i^{MT}(t)$ towards E^{inh} with a strength $G_i^{inh}(t)$.
- The third term of equation (7.3) will drive $u_i^{MT}(t)$ towards the resting potential E^L with a constant strength given by g^L.

The constant δ, typically 30ms, is the delay associated to the inhibitory effect.

The MT neuron i is part of a neural network where the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained by pooling the activity of all the pre-synaptic neurons connected to it. Each MT neuron has a receptive field obtained from the convergence of

[2]The *conductance-based neuron* model is a classical neuron representation. Another models are also possible but their respective differences and performances are out of the scope of this thesis.

pre-synaptic afferent V1 complex cells (see Figure 5.10). The excitatory inputs forming $G_i^{exc}(t)$ are related with the activation of the classical receptive field (CRF) of the MT neuron; whereas $G_i^{inh}(t)$ afferent are the cells forming the surround interactions that could modulate or not the response of the CRF (Xiao et al. (1997b, 1995)) (see Figure 7.2). The surround does not elicit responses by itself, it needs the CRF activation to be considered. According to this, the total input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ of the post-synaptic neuron i are defined by

$$G_i^{exc}(t) = \max\left(0, \sum_{j\in\Omega_i} w_{ij} r_j^{V1} - \sum_{j\in\Omega_i'} w_{ij} r_j^{V1}\right),$$
$$G_i^{inh}(t) = \begin{cases} \sum_{j\in\Phi_i} w_{ij} r_j^{V1} & \text{if } G_i^{exc}(t) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{7.4}$$

where

$$\Omega_i = \{j \in \text{CRF} \mid \varphi_{ij} < \pi/2\}, \tag{7.5}$$
$$\Omega_i' = \{j \in \text{CRF} \mid \varphi_{ij} > \pi/2\}, \tag{7.6}$$
$$\Phi_i = \{j \in \text{Surround} \mid \varphi_{ij} < \pi/2\}, \tag{7.7}$$

and where the connection weight w_{ij} is the efficacy of the synapse from neuron j to neuron i, which is proportional to the angle φ_{ij} between the two preferred motion direction-selectivity of the V1 and MT cell. It is important to remark that the values of the conductances will be always greater or equal to zero, and their positive or negative contribution to $u_i^{MT}(t)$ is due to the values of E^{exc} and E^{inh}.

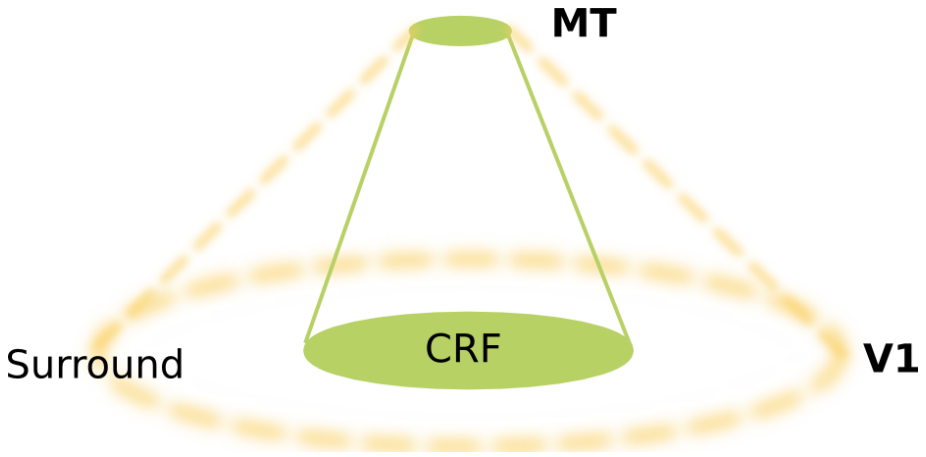


Figure 7.2: MT center-surround construction from V1 neurons. V1 neurons residing inside the classical receptive field (CRF) of the ith MT neuron form part of the excitatory conductance $G_i^{exc}(t)$. Analogously, V1 neurons lying inside the surround of the ith MT neuron form part of the inhibitory conductance $G_i^{inh}(t)$.

About the surround definition in equation (7.7), in Section 5.2.2 we introduced the different center-surround interactions as described by neurophysiology. In this chapter, we implemented this variety of center-surround interactions following the

description given in Figure 5.14, which consists in four different configurations: only CRF, symmetric surround, bilateral asymmetric surround and unilateral symmetric surround.

Remark: *Speed coding relies on complex and unclear mechanisms. Many studies on MT focused on motion direction selectivity (DS), but very few on speed selectivity (see, e.g., Priebe et al. (2003); Perrone and Thiele (2001); Liu and Newsome (2003)). Here we only considered the motion direction and not the motion speed, as can be seen in* (7.4)*: Our MT cells pool V1 cells just considering their motion DS, and not their spatiotemporal tuning. However, note that it could be also possible to pool differently V1 cells in order to extract some speed information, as proposed for example in Simoncelli and Heeger (1998); Grzywacz and Yuille (1990); Perrone (2004). As a result, one could obtain a velocity field qualitatively similar to an optical flow (i.e., one velocity per position).* ■

7.2 TOWARDS HUMAN ACTION RECOGNITION

7.2.1 Supervised classification

Until now, we described a V1-MT model which is inspired by some biological findings. But, how could we use the output of MT neurons in a real application such as human action recognition?

The output of MT neurons will be here used to define feature vectors representing the motion information of the input stimulus, in our case a real video sequence. The feature vector definition has no biological inspiration and it represents the correspondence between the input space (here the space of sequences) and a feature space.

Using the feature vectors obtained from MT cells' output, we considered the simpler case of *supervised* classification which means that for some inputs, the class is known (training set). Then, considering a new sequence to be analyzed, we will estimate the corresponding feature vector and find the best class with a classifier.

7.2.2 Mean Motion Map: Definition of feature vector and distance

In this section, we define the feature vectors as *mean motion maps*, which represent averaged MT cells' activity inside a temporal window, as well as a distance to compare different *mean motion maps*.

At time t, given a video stream $L(\mathbf{x}, t)$ between $[t - \Delta t, t]$, we define the feature vector (from now on called *mean motion map*) as the vector which represents the average membrane potential of the MT neurons inside a sliding temporal window $[t - \triangle t, t]$:

$$H_I(t, \triangle t) = \left\{\gamma_j^L(t, \triangle t)\right\}_{j=1,\ldots,N_l \times N_c}, \tag{7.8}$$

with $\gamma_j^L(t,\triangle t) = \frac{1}{\Delta t}\int_{t-\triangle t}^{t} u_j^{MT}(s)ds$, and where N_l is the number of MT layers and N_c is the number of MT cells per layer. This procedure is summarized in Figure 7.3.

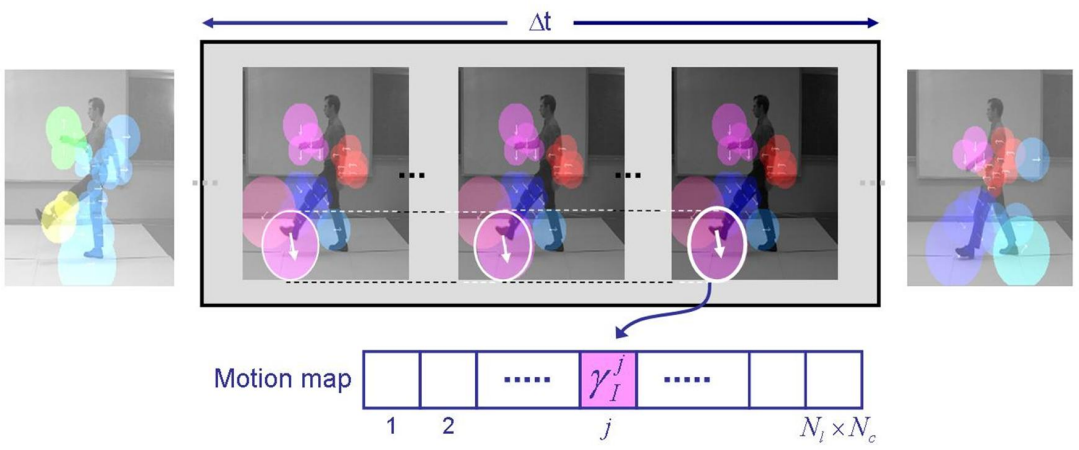


Figure 7.3: *Mean motion map* definition diagram. The membrane potential of the ith MT cell, $r_i^{MT}(t)$, is averaged inside a sliding temporal window of length Δt. The averaged membrane potential $\gamma_i^I(t,\Delta t)$ fills the ith position of the *mean motion map* of length $N_l \times N_c$.

One interesting aspect of the *mean motion map* defined in (7.8) is its invariance to the sequence length and starting point (for Δt high enough depending on the scene). It also includes information regarding the temporal evolution of the activation of MT cells, respecting the causality in the order of events. Besides, the use of a sliding temporal window allows us to include action changes inside the sequence.

Now, in order to evaluate the similarities between two mean motion maps $H_I(t,\triangle t)$ and $H_J(t',\triangle t')$, we propose the following discrimination measure:

$$\mathcal{D}(H_I(t,\triangle t), H_J(t',\triangle t')) = \frac{1}{N_l\, N_c}\sum_{l=1}^{N_l N_c}\frac{(\gamma_l^I(t,\triangle t) - \gamma_l^J(t',\triangle t'))^2}{\gamma_l^J(t,\triangle t) + \gamma_l^J(t',\triangle t')}. \tag{7.9}$$

This measure refers to the *triangular discrimination* introduced by Topsoe (2000). Other measures derived from statistics, such as *Kullback-Leiber* (KL) could also be used. The *Kullback-Leiber* measure was also tested showing no significant improvements. Note that (7.9) and the motion representation (7.8) can be seen as an extension of Zelnik-Manor and Irani (2006), where a similar measure was proposed by the authors to measure the distance between the empirical distributions of each sequence.

7.3 Experiments

7.3.1 Basic validations

Before considering the human action recognition application, the model was tested with simple motion sequences normally used in neurophysiology and psychophysics,

such as, drifting gratings and barberpoles. The results shown in Figure 7.4 and Figure 7.5 are polar diagrams with the outputs of V1 and MT populations. V1 population is tuned to 12 different spatial orientations and 9 different spatiotemporal frequencies. Neurons with the same spatial orientation tuning were grouped and its normalized activity is shown in the respective polar diagrams. MT population is formed by 8 neurons tuned to 8 different spatial orientations. The MT cells considered here have no surround interactions, which means that only the activation of the CRF is taken into account.

See later: *More validations using classical psychophysical stimuli, such as barberpoles and plaids, will be shown in Chapter 9.* ■

We also tested the model on natural sequences. For example, we show in Figure 7.6 the outputs MT neurons for a sequence from the Weizmann database (*jumping-jack denis*). The outputs of the MT neurons most activated are coded by colors following the orientation code shown at the top of Figure 7.6. As we can see, the activation of MT neurons follows the performance of the action, in this case, jumping-jack.

7.3.2 Implementation detail for human action recognition

Input stimuli: are natural image sequence where a single action is being performed. The luminosity and contrast were normalized and the images were resized to 210×210 pixels. The person performing the action was tracked and the video were therefore cropped in order to have the action in the center of the images. We considered 25 frames per second. Samples of the video used in our system are shown in Figure 7.7.

V1 settings: Given V1 cells modeled by (5.9), we consider 9 layers of V1 cells. Each layer is built with V1 cells tuned with the same spatiotemporal frequency and 8 different orientations. The 9 layers of V1 cells are distributed in the frequency space in order to tile the whole space of interest (maximal spatial frequency of 0.5 pixels/sec and a maximal temporal frequency of 12 cycles/sec). The centers of the receptive fields are distributed according to a radial log-polar scheme with a foveal uniform zone. The limit between the two regions is given by the radius of the V1 fovea R_0 (80 pixels). The cells with an eccentricity less than R_0 have an homogeneous density and receptive fields size. The cells with an eccentricity greater than R_0 have a density and a receptive field size depending on its eccentricity, giving a total of 4473 cells per layer.

MT settings: Similarly to V1, MT cells are also distributed in a log-polar architecture, but in this case R_0 is 40 pixels giving a total of 144 cells per layer. Different layers of MT cells conform our model. Four different surround interactions were used

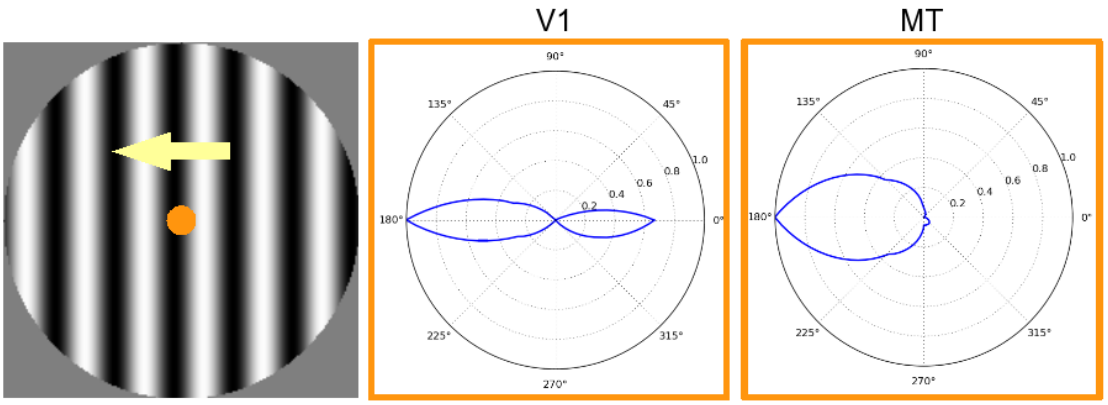


Figure 7.4: Response of a population of V1 and MT neurons to a drifting grating as input stimulus. The grating drifts in the direction of 180°. V1 population is spatially located at the center of the sequence. V1 population is formed with 9 different spatiotemporal frequencies and 12 different spatial frequencies. V1 cells sharing the same spatial orientation were grouped and their normalized activation for each spatial direction are displayed in their respective polar diagram. MT population is also spatially located at the center of the sequence, and it is tuned for eight different spatial orientations. The normalized activation of the MT neurons are displayed in its respective polar graph

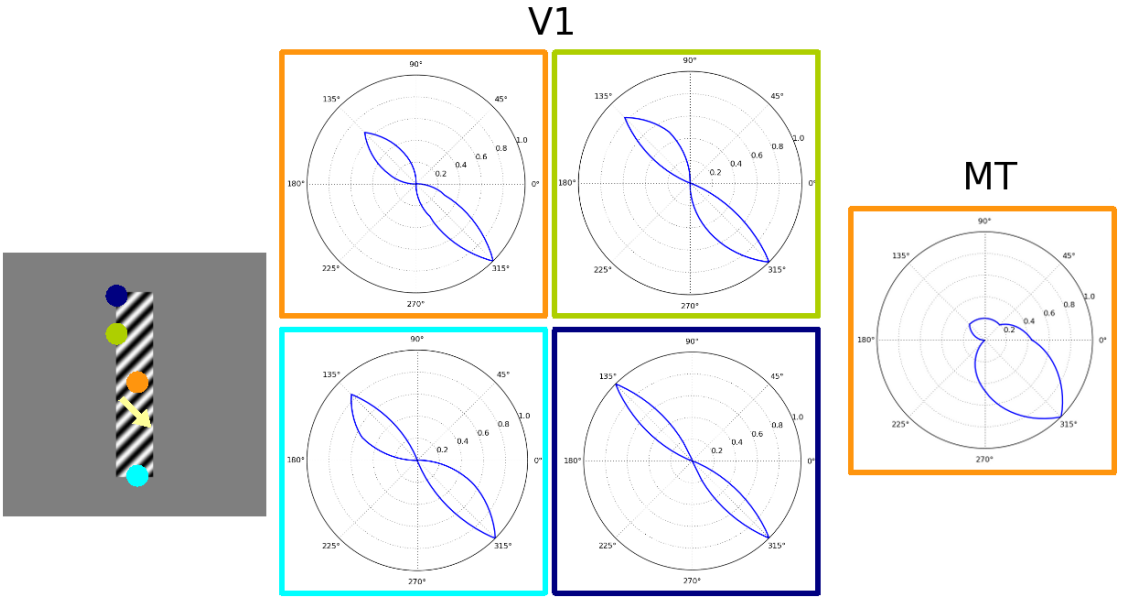


Figure 7.5: Response of a population of V1 and MT neurons to a barberpole of aspect ratio 5:1. The barberpole drifts in the direction of 315°. V1 neurons were placed at the terminators and center of the barberpole, showing for each case, the effect of the respective border in the activation of V1 and MT neurons.

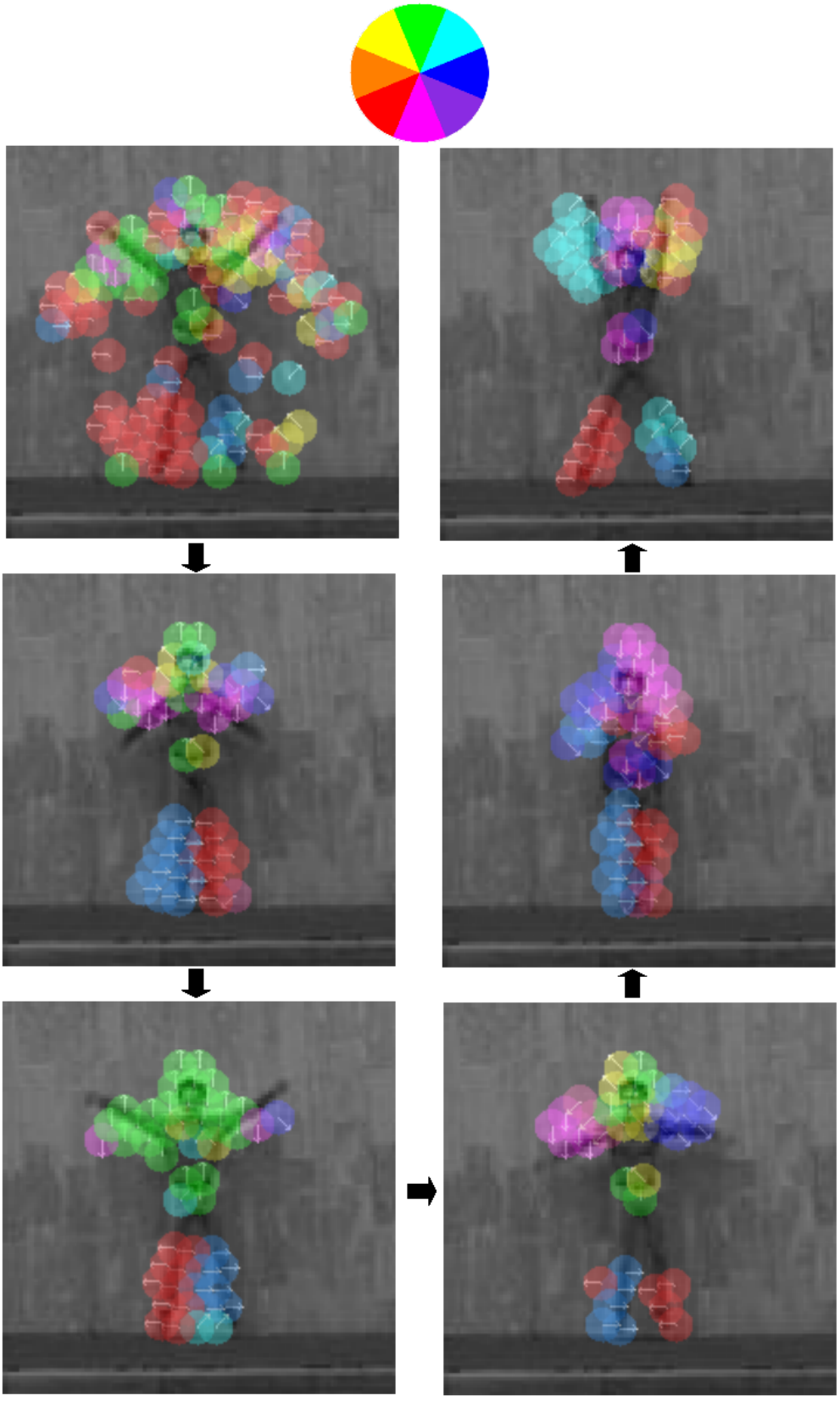


Figure 7.6: Evolution of MT cells along time for a natural video from Weizmann database. The video is a person performing the *jumping-jack* action, and the snapshots at different times show the evolution of the most 20 activated MT cells. The color of each MT cell follows the orientation color code shown at the top.

in the MT construction (see Fig. 3.9). Each layer, with a certain surround interaction, has 8 different directions.

7.3.3 Experimental Protocol

In order to evaluate the performance of our algorithm, we used the Weizmann database[3]. This database contains 9 different samples of different people doing 9 actions: bending (*bend*), jumping jack (*jack*), jumping forward on two legs (*jump*), jumping in place on two legs (*pjump*), running (*run*), galloping sideways (*side*), walking (*walk*), waving one hand (*wave1*) and waving two hands (*wave2*). The number of frames per sequence is variable and depends on the action. A representative frame of each action is shown in Figure 7.7.

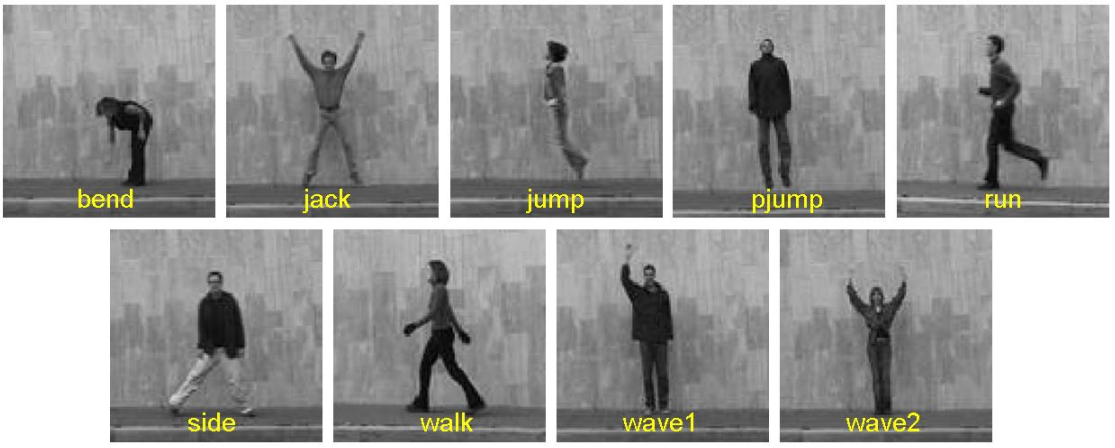


Figure 7.7: Sample frames of each of the nine actions conforming the Weizmann database. The actions are: bending (bend), jumping-jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), galloping-sideways (side), walking (walk), waving-one-hand (wave1) and waving-two-hands (wave2).

We selected the actions of 4 or 6 (as in Jhuang et al. (2007)) random subjects as training set (total of 36 or 64 sequences, respectively) and use the remaining 5 or 3 subjects for the test set (45 or 27 sequences, respectively). All the mean motion maps of the training set were obtained and stored in a data container.

We used a standard classifier[4] defined as follows: When a new input sequence belonging to the test set is presented to the system, the mean motion map is calculated (with Δt covering here all the sequence) and it is compared using (7.9) to all mean motion maps stored in the training set. The class of the sequence with the shortest distance is assigned as the match class.

The experiments were done considering every possible selection of 4 or 6 subjects,

[3] `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`

[4] Note that we repeated the experiments with a standard SVM classifier but we did not get significant improvements in the recognition performance.

giving a total of 126 or 84 experiments. As output we obtained histograms showing the frequency of the recognition error rates.

7.3.4 Results

In order to quantify the influence of the information coded by center-surround interactions, we did the experiments with the different configurations shown in Figure 3.9. The cells were combined in order to create three different mean motion maps: just considering the CRF, CRF plus the isotropic surround interaction, and finally considering all the cells described in Figure 3.9, i.e., with isotropic and anisotropic surround interactions.

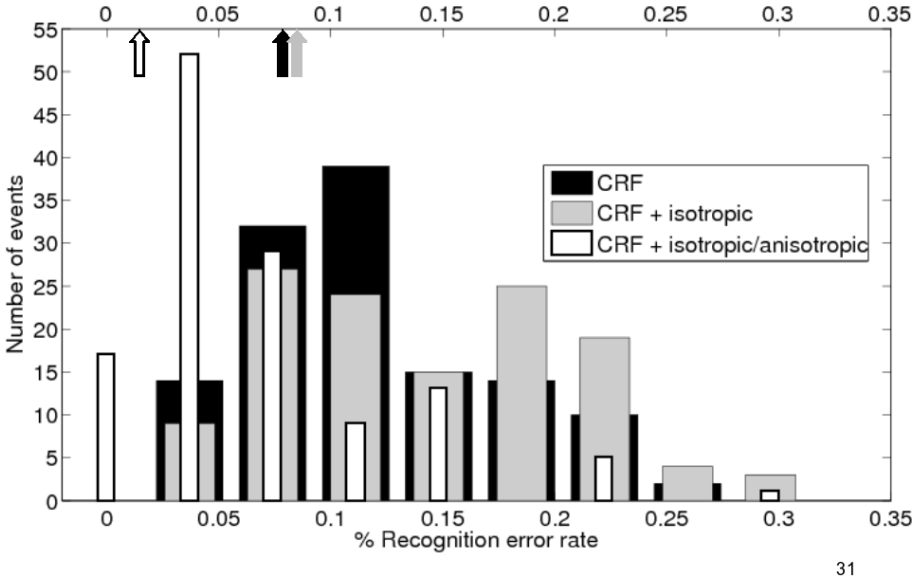
The recognition performance has a strong variability depending on the sequences used to define the training set. This is summarized in the histograms shown in Figures 7.8 and 7.9. From the information contained in the histograms, we can observe that:

- The case of $g^L > 0$ significantly improves the performance of our system, mainly because the constant leak term attracts the membrane potential of the cell to its resting value ($E^L = 0$), avoiding possible saturation.
- The case $g^L = 0$, the effect of inhibitory surrounds (either isotropic or anisotropic) is stronger than the case of $g^L = 0.25$. The explanation is that the inhibitory surround is the only mechanism to reduce the activation of the cell. Maybe this effect can be compensated in the case of $g^L = 0.25$ by adding more relevance to the response of the cells with inhibitory surround. The case $g^L = 0$ converts the leak conductance into a conductance which fully depends on the response of the inhibitory surrounds.

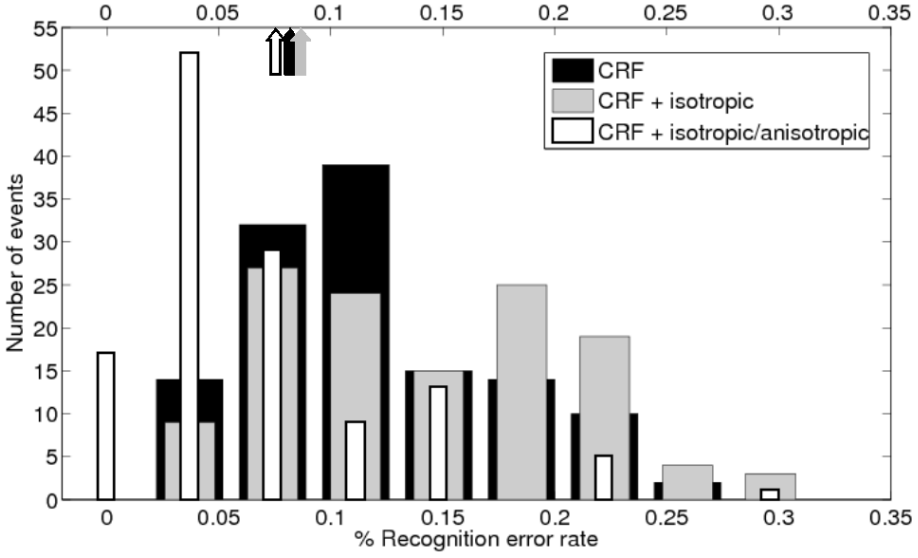
In the case where 6 random subjects were taken to construct the training set, we compared our results with Jhuang et al. (2007). As previously mentioned, we estimated the performance of our approach based on all the possible 84 combinations, and not only on 5 random trials (as in Jhuang et al. (2007)). In Figure 7.10, we show the histogram with the different recognition error rates obtained with our approach using the mean motion maps generated for the CRF and isotropic/anisotropic surround interactions cells. The average recognition rate of 98.9% (i.e., mean error rate of 1.1%), which exceeds the results obtained by Jhuang et al. (2007).

Both, Figure 7.8 and 7.8 have a best recognition performance when all the surround geometries described in Figure 5.14 are considered. A significantly improvement due to different surround geometries is obtained in the case with no leak, i.e. $g^L = 0$, where the value of the membrane potential of MT neurons is really modulated by the different surround configurations. Our interpretation is that singularities in the velocity field are reflected in this modulation.

To test the robustness of our approach, we considered input sequences with different kinds of perturbations (Figure 7.11): noise (case (2)), legs-occlusion (case (3)) and

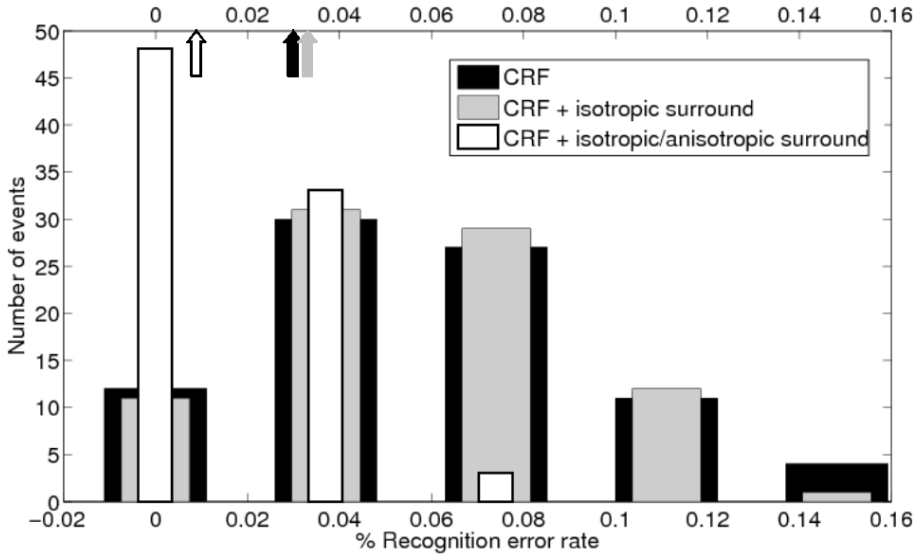


(a)

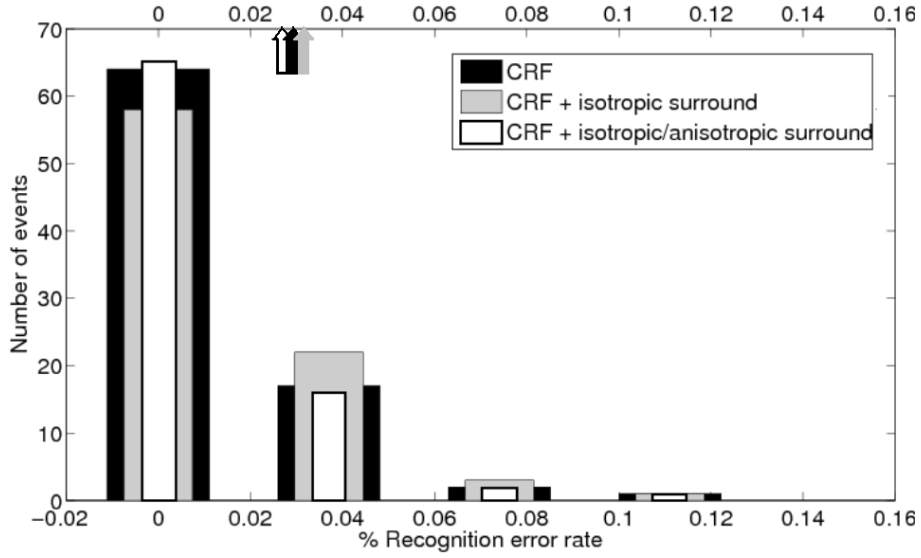


(b)

Figure 7.8: Recognition error rate obtained for Weizmann database. We took all the 126 combinations possibles considering **4 subjects in the training set** (TS). (a) Results obtained for $g^L = 0$. (b) Results obtained for $g^L = 0.25$. All the experiments were performed using the three surround-interactions defined in Figure 3.9: just CRF (black bars), CRF plus isotropic surround suppression (gray bars) and CRF plus isotropic and anisotropic surround suppression (red bars). The mean values for the recognition error rate of each group of cells are shown as arrows at the top of each graph: (a) 15.83%, 19.29%, 9.01%. (b) 7.62%, 8.97%, 7.58%.



(a)



(b)

Figure 7.9: Recognition error rate obtained for Weizmann database. We took all the 84 combinations possibles considering **6 subjects in the training set** (TS). (a) Results obtained for $g^L = 0$. (b) Results obtained for $g^L = 0.25$. All the experiments were performed using the three surround-interactions defined in Figure 3.9: just CRF (black bars), CRF plus isotropic surround suppression (gray bars) and CRF plus isotropic and anisotropic surround suppression (red bars). The mean values for the recognition error rate of each group of cells are shown as arrows at the top of each graph: (a) 3.07%, 3.17%, 0.93%. (b) 5.70%, 7.40%, 5.50%.

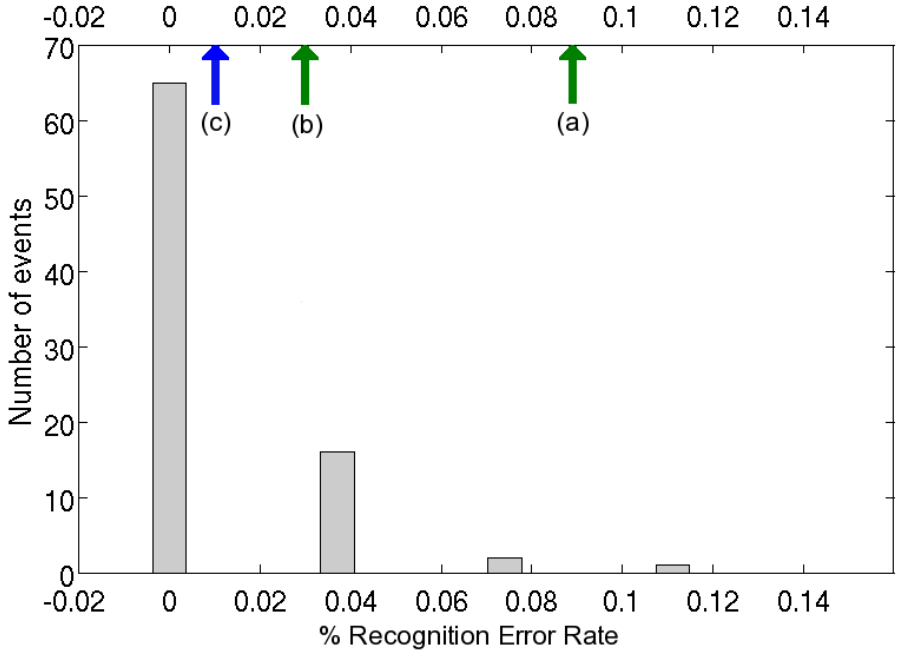


Figure 7.10: Histograms obtained from the recognition error rates of our approach using all the cells defined in Figure 3.9 for Weizmann database and the same experiment protocol used in Jhuang et al. (2007). The gray bars are our histogram obtained for $g^L = 0.25$. (a) Mean recognition error rate obtained by Jhuang et al. (2007) (GrC_2, dense C_2 features): $8.9\%/\pm 5.9$. (b) Mean recognition error rate obtained by Jhuang et al. (2007) (GrC_2, sparse C_2 features): $3.0\%/\pm 3.0$. (c) Mean recognition error rate obtained with our approach: $1.1\%/\pm 2.1$.

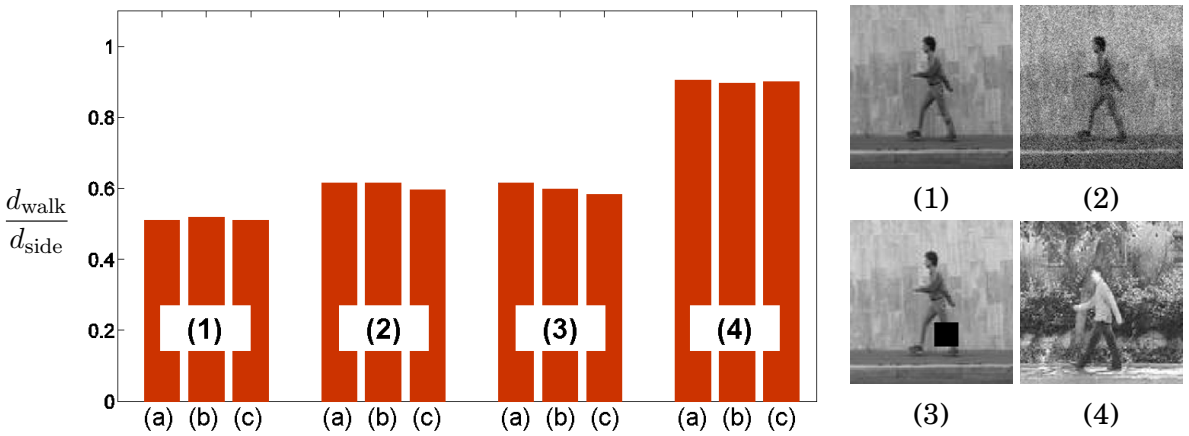


Figure 7.11: Results obtained for the robustness experiments carried out for the three input sequences represented by the snapshots shown for *normal-walker* (1), *noisy* sequence (2), *legs-occluded* sequence (3) and *moving-background* sequence (4). In all the cases the recognition was correctly performed as *walk* and the second closest distance was to the class *side*. The red bars indicate the ratio between the distance to *walk* class and the distance to *side* class (d_{walk}/d_{side}). The experiments were done for the three configurations of surround-suppression: (a) just CRF, (b) CRF with isotropic surround and (c) CRF with isotropic/anisotropic surround ($g^L = 0.25$).

moving textured background (case (4)). Both *noisy* and *legs-occluded* sequences were created starting from the sequence shown in Figure 7.11(1), which was extracted from the training set for the robustness experiments. The *legs-occluded* sequence was created placing a black box on the original sequence before the centered cropping. The *noisy* sequence was created adding Gaussian noise with a variance of ± 30 (for image luminosity varying between 0 and 255). The *moving-background* sequence was taken from Blank et al. (2005). A graph with the ratio between the shortest distance to *walk* class and the distance to the second closest class (*side* for the all the cases) is shown in Figure 7.11. For the original sequence and the three modified input sequences the recognition was correctly performed as *walk*. Also, the inclusion of the anisotropic surround interaction makes the model slightly less sensitive to occlusions or noise.

CHAPTER 8

SPIKING MODEL IMPLEMENTATION

"Spike is the code"
– William Bialek (1997)

Contents

OVERVIEW

In this chapter we define a spiking version of the V1-MT core architecture proposed in Chapter 5, following the same question: can this bio-inspired model be used in human action recognition?

The spiking neurons are modeled as a conductance-based integrate-and-fire neurons. The spike generation mechanism acts as an output nonlinearity that models some of the nonlinearities found in real V1 and MT neurons.

The performance of this architecture is tested with the Weizmann database[1]. The spike trains obtained by MT neurons are processed in two different manners in order to define two motion maps: *mean motion map* and *synchrony motion map*. We evaluate the performance of these two motion maps in the human action recognition task.

Contributions of this chapter

1. Implementation of a spiking V1-MT feedforward architecture for human action recognition.

2. Two complementary interpretations of the neural code are considered to classify actions: mean firing rate of each neuron and synchrony between pairs of neurons.

3. Study of the role of center-surround diversity in MT cells for human action recognition performance.

Keywords: human action recognition, motion maps, mean motion map, synchrony motion map, spikes, spike train analysis, neural code, spike train synchronization, mean firing rate.

Organization of this chapter:

Section 8.1 describes the background for spiking neurons and spike train analysis. Section 8.2 describes the specific implementation for V1 and MT neurons. Section 8.3 show how the output of MT neurons can be used to define two different motion maps. Section 8.4 defines the settings for V1 and MT neurons, the experimental protocol and the human action recognition performance through different measures: recognition error rate, confusion matrices and robustness.

[1] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

8.1 SOME SPIKING BACKGROUND

8.1.1 Introduction

The output of a spiking neural network is a set of events, called spikes, defined by their occurrence times, up to some precision. Spikes represent the way that the nervous system choose to encode and transmit the information. But decoding this information, that is understanding the neural code, remains an open question in the neuroscience community.

There are several hypotheses on how neural code is formed, but there is a consensus on the fact that rate, i.e., the average spiking activity, is certainly not the only characteristic analyzed by the nervous system to interpret spike trains (see, e.g., some early ideas in Perkel and Bullock (1968)). Let us discuss two additional examples.

What about the rank?

For example, rank order coding could explain our performance in ultra-fast categorization. In Thorpe et al. (1996), the authors show that the classification of static images can be performed by the visual cortex within very short latencies: 150 ms and even faster. However, if one consider latency times of the visual stream (Nowak and Bullier (1997)), such timings can only be explained by a specific architecture and efficient transmission mechanisms. As an explanation to the extraordinary performance of fast recognition, rank order coding was introduced (Thorpe (1990); Gautrais and Thorpe (1998)): which means, to interpret the neural code by considering the relative order of spiking times. The idea is that most highly excited neurons fire in average more but also sooner. With this idea of rank order coding, the authors in fact developed a complete theory of information processing in the brain by successive waves of spikes (VanRullen and Thorpe (2002)). Interestingly, the information carried by this first wave has been confirmed by some recent experiments in Gollisch and Meister (2008), where the authors showed that certain retinal ganglion cells encode the spatial structure of a briefly presented image with the relative timing of their first spikes.

What about synchronies and correlations?

Another example of relevant spike train characteristics could be synchronies and correlations. The binding-by-synchronization hypothesis holds that neurons that respond to features of one object fire at the same time, but neurons responding to features of different objects do not necessarily. In vision, neuronal synchrony could thereby bind together all the features of one object and segregate them from features of other objects and the background. Several studies have supported this hypothesis

by showing that synchrony between neuronal responses to the same perceptual object is stronger than synchrony between responses to different objects. Among the numerous observations in this direction, let us mention e.g., Neuenschwander et al. (1999); Fries et al. (2001); Grammont and Riehle (2003) and Biederlack et al. (2006).[2]

But, what about spikes in real vision applications?

Up to our knowledge, there are very few attempts to use spikes in real applications. Moreover, existing work concerns only static images. For example, let us mention two contributions about image recognition (see, e.g., Thorpe (2002) as an application of rank order coding) or image segmentation (see, e.g., Wang and Terman (1995) modeled by oscillator networks), which refer respectively to the two characteristics mentioned above: rank and synchronies.

But analyzing spikes means being able to correctly generate them, which is a difficult issue. At the retina level, some models exist, such as, Thorpe (2002) and Wohrer and Kornprobst (2009) with different degrees of plausibility. When we go deeper in the visual system, we require even more simplifications since it is not possible to render the complexity of all the successive areas and neural diversity. Here, we propose a simplified spiking model of the V1/MT architecture with one goal: Can the spiking output be exploited in order to extract some content like the action taking place?[3]

8.1.2 From spikes to spike trains

The elementary units of the central nervous system are neurons. Neurons are highly connected to each other forming networks of spiking neurons. The neurons collect signals from other neurons connected to it (presynaptic neurons), do some non-linear processing, and if the total input exceeds a threshold, an output signal is generated.

The output signal generated by the neuron is what is known as *spike* or *action-potential*: it is a short electrical pulse that can be physically measured and has an amplitude of about 100mV and a typical duration of 1-2ms. A chain of spikes emitted by one neuron is called *spike train*. The neural code corresponds to the pattern of neuronal impulses (see also Gerstner and Kistler (2002)).

Although spikes can have different amplitudes, durations or shapes they are typically treated as discrete events. By discrete events, we mean that in order to describe

[2]Note that the link between synchrony and segmentation is still controversial. Results could sometimes be explained by other mechanisms taking over the segmentation by synchrony (see, e.g., Roelfsema et al. (2004)).

[3]In spite of these numerous hypothesis, "decoding" the neural code remains an open question in neuroscience (Victor and Purpura (1996); Rieke et al. (1997); Fellous et al. (2004)), which is far beyond the scope of this work. Different metrics or weaker similarity measures between two spike trains have also been proposed (see Cessac et al. (2008) for a review).

a spike train, one only needs to know the succession of emission times:

$$\mathcal{F}_i = \{\ldots, t_i^n, \ldots\}, \text{ with } t_i^1 < t_i^2 < \ldots < t_i^n < \ldots, \tag{8.1}$$

where t_i^n corresponds to the nth spike of the neuron of index i.

8.1.3 Interpretations of the neural code

The set of all spikes from a set of neurons in a period of time is generally represented in a graph called *raster plot*, as illustrated in Figure 8.1.

The question is to know how this pattern of neuronal impulses is analyzed by the nervous system. The most simple and intuitive is to estimate the mean firing rate over time, which is the average number of spikes inside a temporal window (rate coding). But what makes the richness of such representation is the many other ways to analyze it. For example, one thought about rate coding over several trials or over population of neurons, time to first spike, phase, synchronization and correlations, interspike intervals distribution, repetition of temporal patterns, etc.

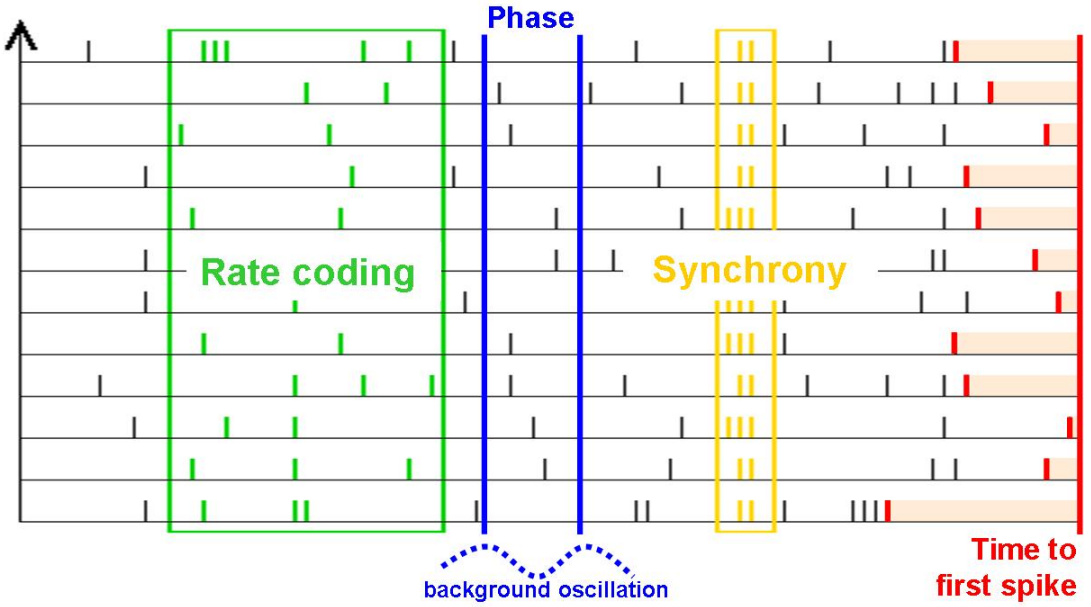


Figure 8.1: Example of a raster plot and illustration of some different methods to analyze the neural code (see text for more details). Each horizontal line can be interpreted as an axon in which we see spikes traveling (from left to right).

8.1.4 Spike train analysis: Example of two measures

Given spike trains as output of a spiking network of neurons, let us propose in this section the two measures we chose to describe its activity: the mean firing rate of a spike train and a synchrony measure between pairs of spike trains. These two

measures will be then directly used in the action recognition application described in Section 8.4.

Remark: *: Note that we do not consider high-level statistics of spike trains (Rieke et al. (1997)), since this requires large ergodic spike sequences, whereas we are interested here in recognition tasks from non-stationary spike trains generated by some dynamic input. Also, we do not considered spike-train metrics in the strict sense (Victor and Purpura (1996)), since we do not have enough knowledge from the biology to predict the firing times in a deterministic way. For the same reason, we do not compare, here, spike patterns (Fellous et al. (2004)). These aspects will be perspectives of this thesis.* ■

Measure 1: Mean firing rate of a neuron

Let us consider a spiking neuron i. The spike train $\mathcal{F}_i$ associated to this neuron is defined in (8.1). We defined the *windowed firing rate* $\gamma_i(\cdot)$ by

$$\gamma_i(t, \Delta t) = \frac{\eta_i(t - \Delta t, t)}{\Delta t}, \tag{8.2}$$

where $\eta_i(\cdot)$ counts the number of spikes emitted by neuron i inside the sliding time window $[t - \Delta t, t]$ (see Figure 8.2 and, e.g., Dayan and Abbott (2001)).

Measure 2: Synchrony between two spike trains

Let us consider the recent approach proposed by Kreuz et al. (2007) to estimate the similarity between two spike trains, as a measure of synchrony. The authors proposed to compute first the interspike interval (ISI) instead of the spike as a basic element of comparison. The use of ISI has the main advantage to be parameter-free and self-adaptive, so that there is no need to fix a time scale beforehand ("binless") or to fit any parameter.

So, for the neuron i the ISI representation $ISI_i(t)$ is given by

$$ISI_i(t) = \min(t_i^{(f)} \mid t_i^{(f)} > t) - \max(t_i^{(f)} \mid t_i^{(f)} < t), \tag{8.3}$$

for $t_i^{(f)} < t$. Considering the ISI representation of two neurons i and j, the next step is to calculate the ratio $R_{ij}(t)$ defined as

$$R_{ij}(t) = \begin{cases} \frac{ISI_i(t)}{ISI_j(t)} - 1 & \text{if } ISI_i(t) \leq ISI_j(t), \\ -\left(\frac{ISI_j(t)}{ISI_i(t)} - 1\right) & \text{otherwise.} \end{cases} \tag{8.4}$$

$R_{ij}(t)$ will be zero in case of completely synchrony between $ISI_i(t)$ and $ISI_j(t)$. In the cases of a big difference between the two ISI representation, $R_{ij}(t)$ will tend to ± 1. (see Figure 8.3).

Having the ratio $R_{ij}(t)$ it is possible to calculate, for a finite time Δt, a measure of spike train distance $\zeta_{ij}(t; \Delta t)$, which is an estimator of the spike train synchrony

between neurons i and j.

$$\zeta_{ij}(t;\Delta t) = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} |R_{ij}(s)|ds. \tag{8.5}$$

Remark: *: Completely synchrony $\zeta_{ij}(\cdot) = 0$ was assigned for two cells not emitting spikes, while the maximal desynchronization $\zeta_{ij}(\cdot) = 1$ was assigned to the case where only one cell emitted spikes.* ■

8.1.5 Spiking neuron modelization

Many spiking neuron models have been proposed in the literature. They differ by their biological plausibility and their computational efficiency (see Izhikevich (2004) for a review).

The V1 and MT cells will be here modeled as a conductance-driven integrate-and-fire neuron (Wielaard et al. (2001); Destexhe et al. (2003)). Considering a neuron i, defined by its membrane potential $u_i(t)$, the integrate-and-fire equation is given by:

$$\frac{du_i(t)}{dt} = G_i^{exc}(t)(E^{exc} - u_i(t)) + G_i^{inh}(t)(E^{inh} - u_i(t)) + g^L(E^L - u_i(t)) + I_i(t), \tag{8.6}$$

with the spike emission process:

- The neuron i will emit a spike when the normalized membrane potential of the cell $u_i(t)$ reaches the threshold μ, i.e., $u_i(t) = \mu$.
- $u_i(t)$ is then reinitialized to its resting potential E^L.

The typical values for the reverse potentials E^{exc}, E^{inh} and E^L are 0mv, -80mV and -70mV, respectively (see Figure 8.4 for an illustration). The neuron membrane potential $u_i(t)$ will evolve according to inputs through either conductances ($G_i^{exc}(t)$ or $G_i^{inh}(t)$) or external currents ($I_i(t)$). $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the pre-synaptic neurons connected neuron i. The conductance g^L is the passive leaks in the cell's membrane. $I(t)$ is an external input current. Finally, $G_i^{inh}(t)$ is an inhibitory normalized conductance dependent on, e.g., lateral connections or feedbacks from upper layers.

8.2 SPIKING V1-MT ARCHITECTURE

8.2.1 V1 neuron implementation

The response of the V1 complex cell -formed as a combination of the V1 simple cells defined in (5.2)- is analog. To transform the analog response into a spiking response, the cell will be modeled as a conductance-driven integrate-and-fire neuron described in (8.6).

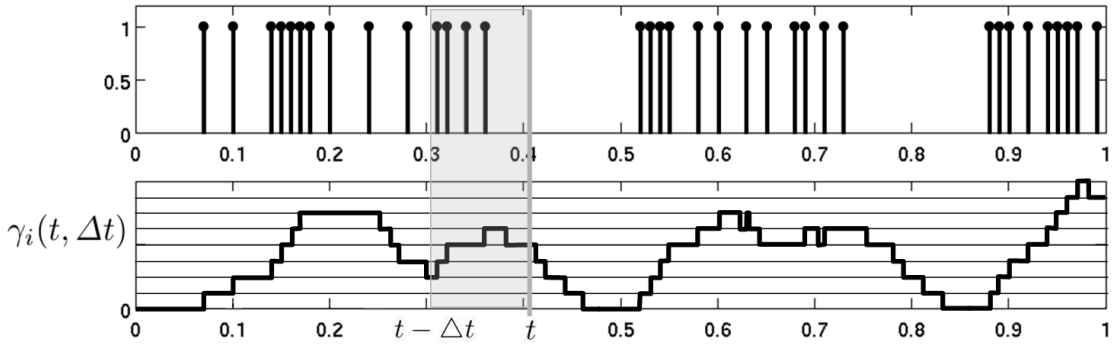


Figure 8.2: Mean firing rate of a spike train

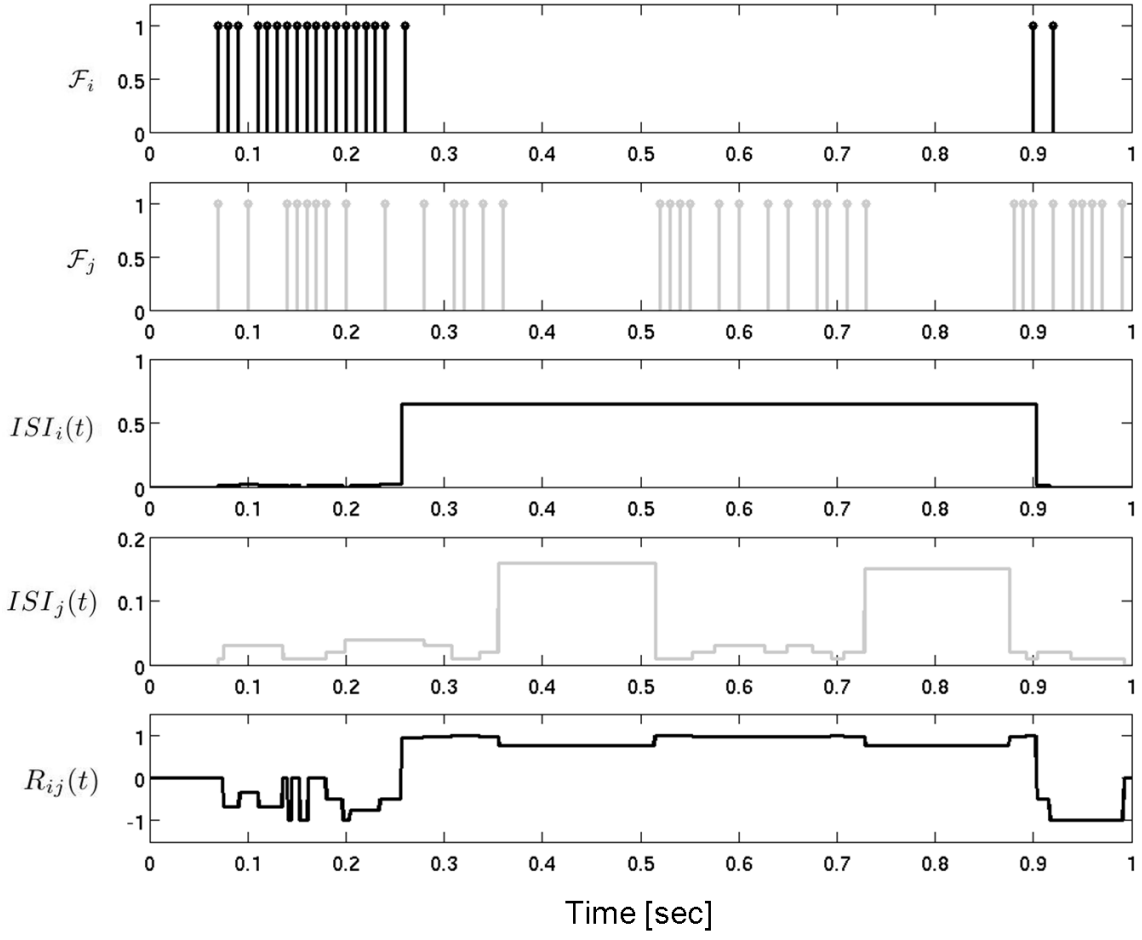


Figure 8.3: Synchrony between the spike trains of a pair of neurons. $\mathcal{F}_i$ and $\mathcal{F}_j$ are the spike trains of MT neurons i and j, respectively. The respective ISI representations defined in (8.3) are shown as $ISI_i(t)$ and $ISI_j(t)$. Finally, the ratio between $ISI_i(t)$ and $ISI_j(t)$ is shown as $R_{ij}(t)$.

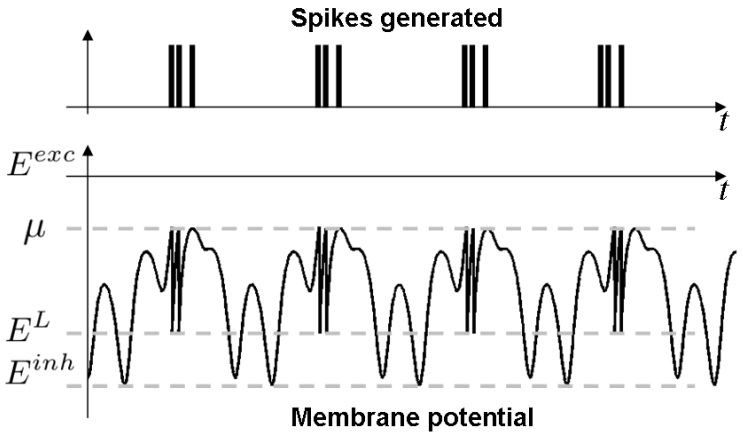


Figure 8.4: Temporal evolution of the membrane potential of a neuron and its corresponding spikes. A spike is generated when the membrane potential exceeds the threshold μ ($E^L < \mu < E^{exc}$). When the spike is emitted membrane potential returns to its resting value E^L.

So, let us consider a spiking V1 complex cell i whose center is located in $\mathbf{x}_i = (x_i, y_i)$ of the visual space. For this neuron, $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the pre-synaptic neurons connected to V1 cells. The external input current $I_i(t)$ is here associated with the analog V1 complex cell response. So, $I_i(t)$ of the ith cell in (8.6) is modeled as

$$I_i(t) = k_{exc} C_i(\mathbf{x}_i, t), \tag{8.7}$$

where k_{exc} is an amplification factor, $C_i(\cdot)$ refers to the complex cell response defined in (5.9). For V1 neuron implementation, the leak conductance, the inhibitory and excitatory conductances of (8.6) are not considered, so that:

$$\begin{aligned} G_i^{inh}(t) &= 0, \\ G_i^{exc}(t) &= 0, \\ g^L &= 0, \end{aligned} \tag{8.8}$$

8.2.2 MT neuron implementation

Similarly, let us model MT cell i as a conductance-driven integrate-and-fire neuron (see equation (8.6)).

Each MT cell has a receptive field made from the convergence of afferent V1 complex cells (see Figure 5.10). Those inputs will be excitatory or inhibitory depending on the characteristic and shape of the corresponding MT receptive fields (Xiao et al. (1997b, 1995)).

The MT neuron i is a part of a spiking network where no external input current is considered ($I_i(t) = 0$) and the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained considering the activity of all the pre-synaptic neurons connected to it (see Figure

5.10 and equation 5.14). For example, if a pre-synaptic neuron j has fired a spike at time $t_j^{(f)}$, this spike reflects an input conductance to the post-synaptic neuron i during a time course $\alpha(t - t_j^{(f)})$. So we have:

$$G_i^{exc}(t) = \sum_{j\in\Omega} w_{ij}^{+} \sum_f \alpha(t - t_j^{(f)}; \tau_s),$$
$$G_i^{inh}(t) = \sum_{j\in\Phi} w_{ij}^{-} \sum_f \alpha(t - t_j^{(f)}; \tau_s), \tag{8.9}$$

where the coefficients w_{ij}^{+} (w_{ij}^{-}) are the efficacy of the positive (negative) synapse from neuron j to neuron i (see Gerstner and Kistler (2002) for more details) and their respective values are defined in equation (5.13). The time course $\alpha(s; \tau_s)$ of the post-synaptic current in (8.9) can be modeled as an exponential decay with time constant τ_s as follows

$$\alpha(s; \tau_s) = \left(\frac{s}{\tau_s}\right) \exp\left(-\frac{s}{\tau_s}\right). \tag{8.10}$$

The domains Ω and Φ, are defined according to the shape of the MT classical receptive field and MT surround, respectively. Inspired by the organization of different center-surround interactions reported by Born (2000) (see more details in Section 3.2.3), we implemented the three types of center-surround interactions defined in Section 5.2.2, specifically in Figure 3.9.

8.3 Towards Human Action Recognition

Similarly to Section 7.2, the output of MT neurons will be used to construct feature vectors representing the motion information of the input stimulus. These feature vectors will be further used in a *supervised* classifier to perform human action recognition.

In this chapter, two different feature vectors are proposed. Those feature vectors: *mean motion map* and *synchrony motion maps*, are based on the spike train measures defined in Section 8.1.4.

8.3.1 Mean Motion Map: Definition of feature vector

The *mean motion map* (equivalent to the *mean motion map* defined in Section 7.2) $H_L(\cdot)$, representing the input stimulus $L(x, y, t)$, is defined by

$$H_L(t, \triangle t) = \left\{\gamma_j^L(t, \triangle t)\right\}_{j=1,\ldots,N_l\times N_c}, \tag{8.11}$$

where N_l is the number of MT layers and N_c is the number of MT cells per layer. Each element γ_j^L with $j = 1, ..., N_l \times N_c$ is the windowed firing rate defined in (8.2) (see Figure 7.3).

The comparison between two mean motion maps $H_L(t, \triangle t)$ and $H_J(t, \triangle t)$, is computed using the distance measure defined in equation (7.9).

8.3.2 Synchrony Motion Map: Definition of feature vector and distance

As it is shown in Section 8.1.4, for each pair of cells $\{i, j\}$ it is possible to obtain a measure of synchrony using $\zeta_{ij}(\cdot)$ defined in (8.5).

Let us consider N_l population of MT cells. For each population, we created a matrix containing the values of $\zeta_{ij}(t; \Delta t)$ obtained to every pair of cells $\{i, j\}$ in the population. The values of $\zeta_{ij}(t; \Delta t)$ were computed inside a sliding time window of size Δt. So, each sequence L will be represented by a *synchrony motion map* $\tilde{H}_L(t, \Delta t)$ defined as

$$\tilde{H}_L(t, \Delta t) = \{\mathrm{D}_k^L(t; \Delta t)\}_{k=1..N_l}, \tag{8.12}$$

where $\mathrm{D}_k^L(\cdot) = \{\zeta_{mn}(\cdot)\}_{m=1..N_c, n=1..N_c}$ is a matrix of $N_c \times N_c$ elements containing the measures $\zeta_{mn}(\cdot)$ between the mth and nth neurons of the kth population of MT cells defined in (8.5). The $\tilde{H}_L$ construction can be summarized in Figure 8.5.

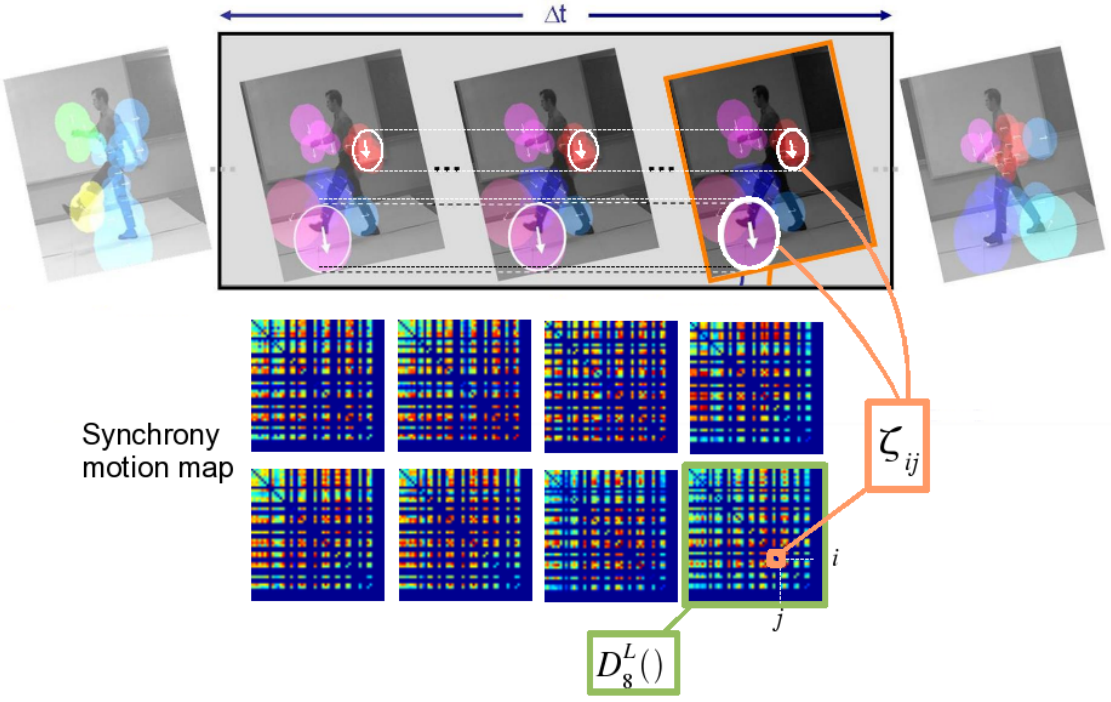


Figure 8.5: Schematic diagram summarizing the *synchrony motion map* construction for the MT neurons located at two different positions of the image. In this case, there are 8 populations of MT neurons ($N_l = 8$), obtaining like this 8 different $\mathrm{D}_k^L(t; \Delta t)$ matrices containing the respective ζ_{ij} values.

The comparison between two synchrony motion maps $\tilde{H}_L(t, \Delta t)$ and $\tilde{H}_J(t', \Delta t')$. is defined by the euclidean distance

$$\tilde{\mathcal{D}}(\tilde{H}_L(t, \Delta t), \tilde{H}_J(t', \Delta t')) = \sqrt{\sum_{N_l} \|\mathbf{D}_k^L - \mathbf{D}_k^J\|^2}. \tag{8.13}$$

Table 8.1: Parameters used for V1 and MT layers.

	V1	MT
Fovea radius	80[pixels]	40[pixels]
Layer radius	100[pixels]	100[pixels]
Cell density in fovea	0.4[cells/pixel]	0.1[cells/pixel]
Eccentricity decay	0.02	0.02
Radius receptive field in fovea	$2\sigma_{V1}$[pixels]	9[pixels]
Number orientations	8	8
Number cells per layer	3302	161

8.4 EXPERIMENTS

8.4.1 Implementation detail for human action recognition

Input stimuli: Same as the stimuli described in Section 7.3.2.

V1 settings: V1 has a total of 72 layers, formed by 8 spatial orientations and 9 different spatiotemporal frequencies, giving a total of 3302 cells per layer. Following the biological result mentioned in Watson and Ahumada (1983) the value of σ_{V1} is $0.5622/f$. The 72 layers of V1 cells are distributed in the frequency space in order to tile the whole space of interest. We considered a maximal spatial frequency of 0.5 pixels/sec and a maximal temporal frequency of 12 cycles/sec.

MT settings: In the case of MT, 8 (1×8 orientations) or 24 (3×8 orientations) layers were used depending on the center-surround configuration defined in Figure 3.9.

General V1 and MT settings are shown in Table 8.1.

8.4.2 Experimental protocol

We ran the experiment using Weizmann[4] database. Weizmann database consists in 9 different subjects performing 9 different actions. A representative frame of each action is shown in Figure 7.7. The number of frames per sequence is variable and the original video streams were resized and centered to have sequences of 210×210 pixels.

The performance of the bio-inspired spiking V1-MT model is here evaluated in the human action recognition application. The system follows the architecture described in Figure 5.1. The outputs of V1 motion detectors feed V1 neurons as an external current. The spike trains generated by V1 neurons feed the MT layers where the

[4] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

activation of each MT neuron depends on the activation of the V1 stage. Figure 8.6 shows the spike trains generated by MT neurons for two different sequences of Weizmann database.

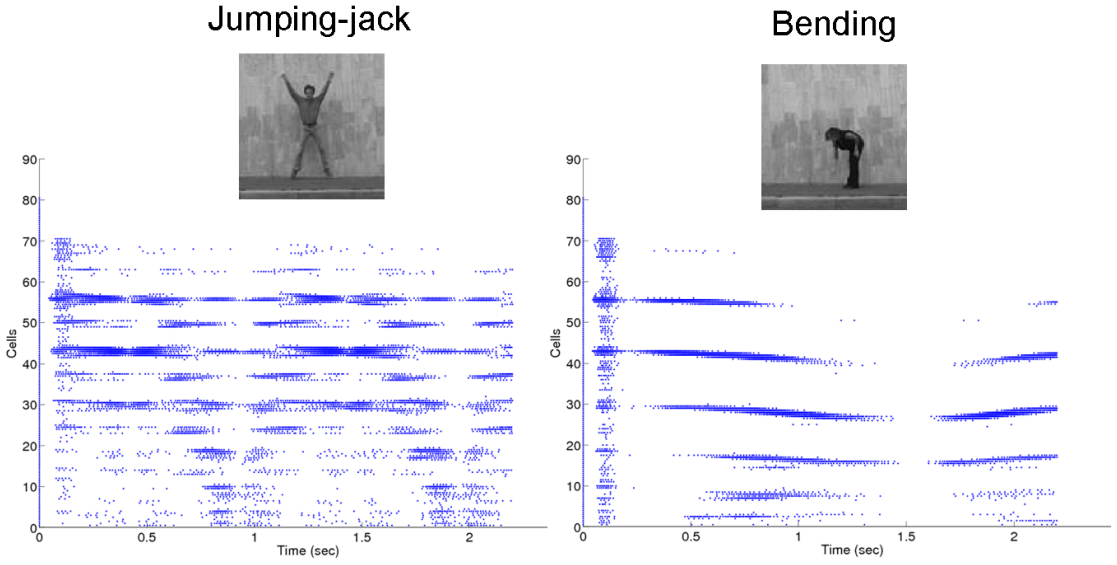


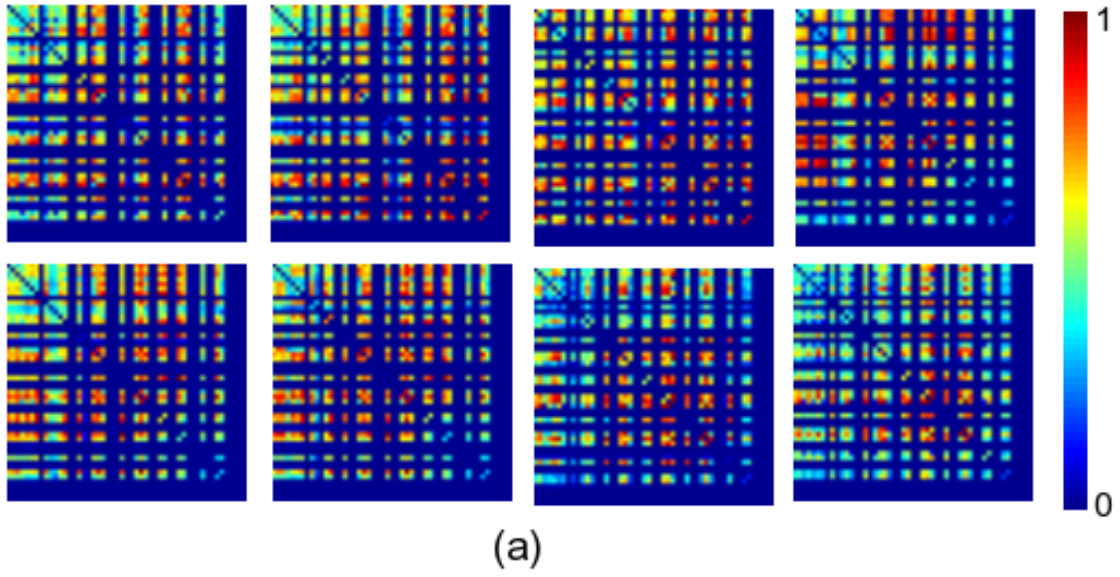
Figure 8.6: Raster plots obtained considering 161 MT cells with only CRF of a given orientation in two different actions: *jumping-jack* and *bending*. Looking at the raster plots obtained, is evident that the information contained into the spike trains is much richer than a simplified mean firing rate. The frame rate is 25 frames per seconds.

To evaluate the recognition performance of our approach using the motion maps defined in Sections 8.3.1 and 8.3.2, we followed a similar experimental protocol than the one proposed by Jhuang et al. (2007). The *mean motion maps* and *synchrony motion maps* of all the 81 sequences forming Weizmann database were calculated, removing in both cases the first 5 frames containing initialization information. A real example showing the synchrony motion maps obtained for two different sequences of the Weizmann database, and for 8 populations of MT neurons ($N_l = 8$) is shown in Figure 8.7.

Similarly to Section 7.3.3, the *training set* was built considering actions of 6 different subjects (6 subjects $\times$ 9 actions = 54 motion maps). The *testing set* was built with the remaining 3 subjects (3 subjects $\times$ 9 actions = 27 motion maps). Unlike Jhuang et al. (2007), we ran all the possible training sets (84) and not only 5 random trials. Each motion map is compared to every motion map in the training set. The match class will be the class associated to the motion map with the lowest distance (according to equation (8.5)).

8.4.3 Results

For each training set, the experiment was performed twice: one time considering 8 layers of MT cells ($N_l = 8$) with the activation of the CRFs for the 8 different



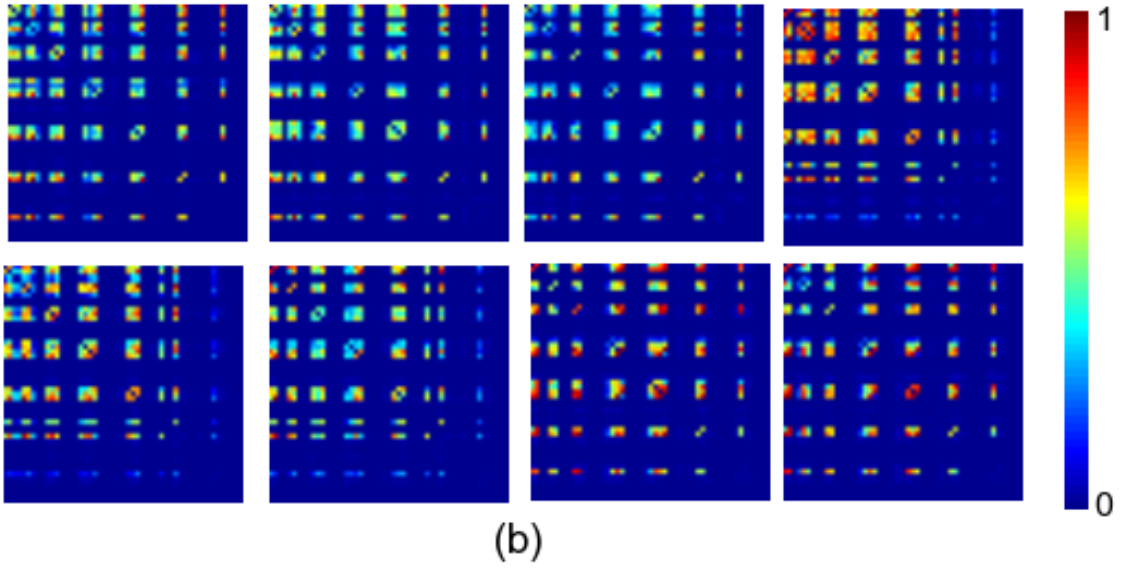


Figure 8.7: Matrices conforming the *synchrony motion maps* defined in (8.12). Each matrix shows the synchronization (see (8.5)) between the spike trains of iso-oriented cells members of the same MT population. (a) Synchrony motion map for a *jumping-jack* sequences of Weizmann database. (b) Synchrony motion map for a *bending* sequence of Weizmann database. One observes significant differences between the synchrony maps of both actions.

Table 8.2: Mean recognition error rates and standard deviation obtained by our approach and by Jhuang et al. (2007).

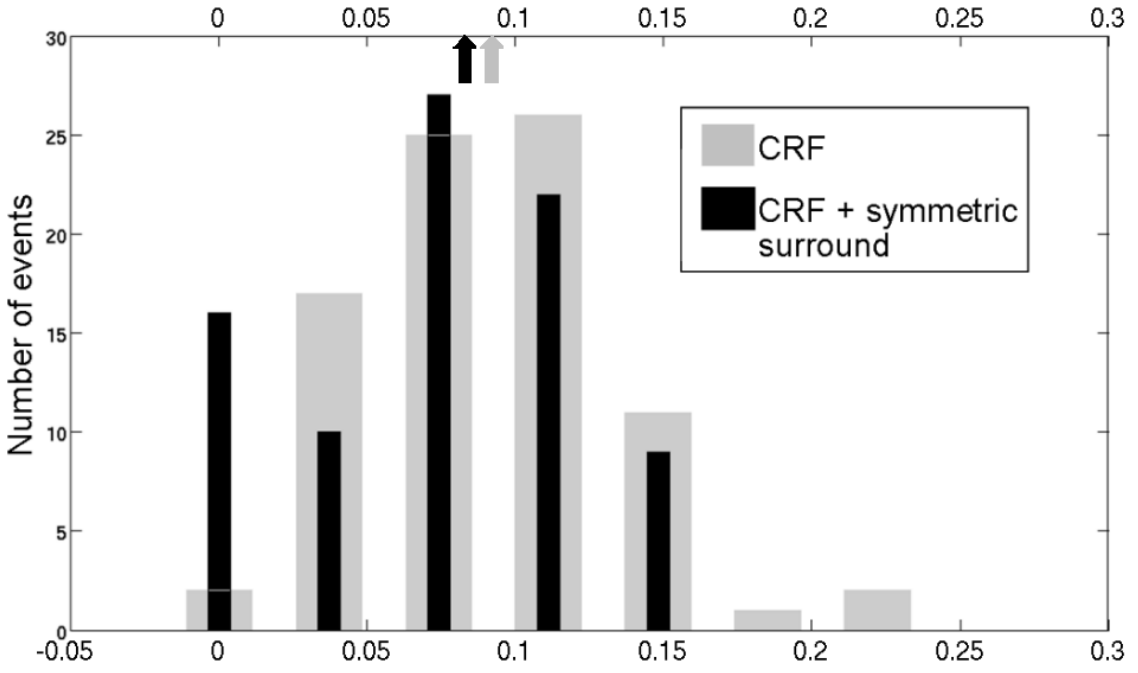
	Mean error rate $\pm$ STD	#trials
Jhuang et al. GrC_2 dense C_2 features	$8.9\%/\pm 5.9$	5
Jhuang et al. GrC_2 dense C_2 features	$3.0\%/\pm 3.0$	5
Mean motion maps CRF	$9.08\% \pm 4.40$	84
Mean motion maps CRF + symmetric surrounds	$7.32\% \pm 4.62$	84
Synchrony motion maps CRF	$13.89\% \pm 4.95$	84
Synchrony motion maps CRF + symmetric surrounds	$7.19\% \pm 5.15$	84

orientations, and a second time with 24 layers of MT cells ($N_l = 24$) using, for each orientation, all the surround interactions shown in Figure 3.9. We constructed a histogram with the different recognition error rates obtained by our approach (see Figure 8.8) using *mean motion maps* and *synchrony motion maps*. As we can see in Figure 8.8, the values have a strong variability and the recognition performance highly depends on the sequences used to construct the training set, reaching in most of the cases 100% of correct recognition.

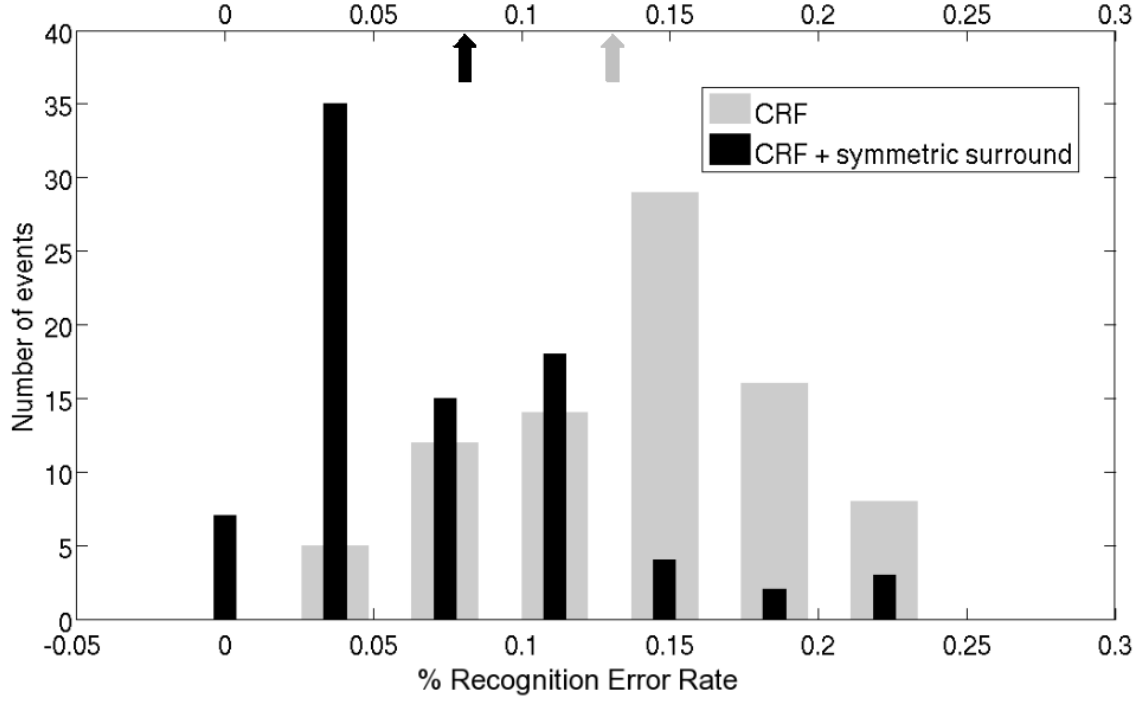
A comparison with the results obtained by Jhuang et al. (2007) is shown in Table 8.2. It is important to remark that our results were obtained using the 84 training sets built with 6 subjects (i.e., all possible combinations) and not only 5 trials as in Jhuang et al. (2007). As it was previously remarked, because of the high variability of classification performance depending on the training set chosen, results in Jhuang et al. (2007) are hard to interpret.

Confusion matrices

In order to have a qualitative comparison between the quality of the human action representation using the two motion maps defined in Section 8.3, we estimated the *confusion matrices* for the 81 sequences conforming Weizmann database (see Figure 8.9). The sequences were grouped according to the action performed (total of 9 actions), and for each pair of actions the mean distance value was obtained. The matrices are 9×9 and they were built using $N_l = 8$ (just MT CRF) and $N_l = 8 \times 3$ (using the three MT center-surround interactions of Figure 3.9). Interestingly, despite of the lower recognition performance of *synchrony motion maps* compared with *mean motion maps*, *synchrony motion maps* better separates the data belonging to different



(a)



(b)

Figure 8.8: Histograms representing the recognition error rates obtained by our approach in Weizmann database, using: MT CRFs (gray bars) and MT center-surround interactions shown in Figure 3.9 (black bars). The results were obtained using the 84 possible training sets built with 6 different subjects. The respective mean values are displayed at the top of each graph (see Table 8.2 for details). (a) Histogram obtained for *mean motion maps* (b) Histogram obtained using *synchrony motion maps*.

classes, specially for actions were only a limited part of the body performs the motion (*waving-one-hand*, *waving-two-hands*, *bending*).

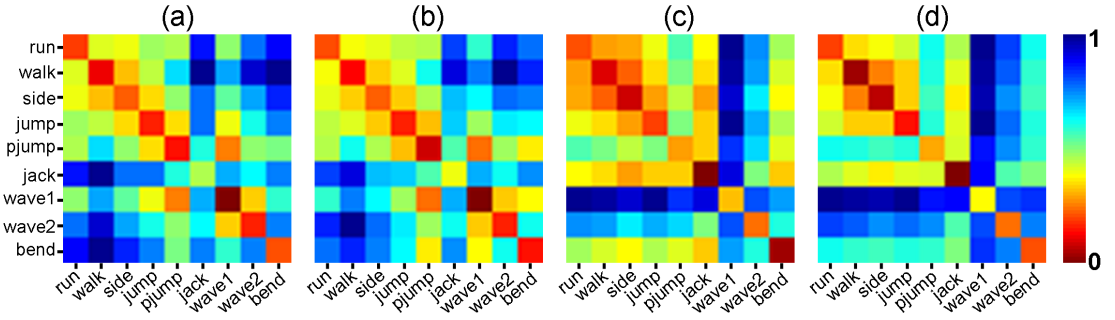


Figure 8.9: Confusion matrices obtained using two different readouts: (a)-(b) *mean motion maps* defined in (8.11) and (c)-(d) *synchrony motion maps* defined in (8.12). We also compare: (a)-(c) considering only MT CRFs and (b)-(d) considering all the MT center-surround interactions defined in Figure 3.9

In order to quantify the inter-class separability we applied a simple statistical analysis (t-student test). Applying the t-student test on the obtained distances matrices we numerically observe for intra-class distances a range of t-value $\in [0.20; 0.26]$ for *mean motion maps* and t-value $\in [0.29; 0.31]$ for *synchrony motion maps*, which in term of probabilities means that the probability to have distances different of zero is $P < 0.60$ and $P < 0.61$, respectively. A significant difference is seen in the inter-class distances, where the range of t-values for *running/all-other-sequences* is t-value $\in [1.40; 2.93]$ (*synchrony motion maps*) and t-value $\in [0.44; 0.55]$ (*mean motion maps*). This can be interpreted, for instance, that for *jumping*/*walking* the distances are different from 0 with a probability of $P < 0.69$ for *mean motion maps* and $P < 0.90$ for *synchrony motion maps*. Although t-test values obtained for *mean motion maps* are numerically higher for inter-class than intra-class distances, it appears that they are not "significantly" higher compared to the ones obtained with the *synchrony motion maps*.

Robustness

To evaluate some kind of robustness of the approach, similarly than Section 7.3.4, we considered input sequences with perturbations. Snapshots of the sequences considered to measure the robustness of the model are shown in Figure 8.10. We considered three kinds of perturbations: *noisy* sequence (Figure 8.10 (**2**)), *legs-occluded* sequence (Figure 8.10 (**3**)) and *moving-background* sequence (Figure 8.10 (**4**)). Both *noisy* and *legs-occluded* sequences were created starting from the sequence shown in Figure 8.10 (**1**), which was extracted from the training set for the robustness experiments. The *legs-occluded* sequence was created placing a black box on the original sequence before the centered cropping. The *noisy* sequence was created adding a Gaussian noise with a variance of ± 30 (for image luminosity varying between 0 and 255). The

Table 8.3: The Null hypothesis rejection probability associated with the t-test values obtained from the distance matrices built using *mean motion maps* and *synchrony motion maps* (case CRF + symmetric surrounds). The corresponding action for each value is the same than the ones shown in Figure 8.9.

Mean motion map

0.59	0.70	0.71	0.69	0.68	0.62	0.67	0.68	0.72
0.70	0.59	0.69	0.68	0.72	0.65	0.70	0.70	0.74
0.71	0.69	0.60	0.66	0.68	0.63	0.68	0.69	0.72
0.69	0.68	0.66	0.60	0.72	0.62	0.68	0.69	0.75
0.68	0.72	0.68	0.72	0.60	0.59	0.64	0.66	0.72
0.62	0.65	0.63	0.62	0.59	0.59	0.61	0.64	0.65
0.67	0.70	0.69	0.68	0.64	0.61	0.58	0.64	0.69
0.68	0.70	0.69	0.69	0.66	0.64	0.64	0.58	0.68
0.72	0.74	0.72	0.75	0.72	0.65	0.69	0.68	0.59

Synchrony motion map

0.61	0.86	0.88	0.90	0.97	0.98	1.00	0.99	0.99
0.86	0.62	0.89	0.90	0.97	0.94	0.99	0.98	0.97
0.88	0.89	0.62	0.86	0.98	0.97	1.00	0.96	0.98
0.90	0.91	0.86	0.62	0.99	0.96	1.00	0.99	0.99
0.97	0.97	0.98	0.99	0.61	0.85	0.93	1.00	0.91
0.98	0.94	0.97	0.96	0.85	0.62	0.96	0.93	0.94
1.00	0.99	1.00	1.00	0.93	0.96	0.60	0.76	0.86
0.99	0.98	0.96	0.99	0.99	0.93	0.75	0.60	0.96
0.99	0.97	0.98	0.99	0.91	0.94	0.86	0.96	0.61

moving-background sequence was taken from Blank et al. (2005). Both, the original sequence and the three modified input sequences, the recognition was correctly performed as *walking* when the *mean motion maps* were used.

The bars of Figure 8.10 represent the ratio between the shortest distance to *walking* (d_{walk}) class and the distance to the second closest class (d_*), which can vary from *galloping-sideways* to *bending* or *jumping-forward-in-two-legs*. Note that in most of the cases the action was correctly recognized as *walking*, giving a ratio $d_{walk}/d_* < 1$. The recognition failed in the case of *synchrony motion maps* (a) who consider only the CRF activation. In those cases the action was always misclassified as *bending* ($d_{walk}/d_{bend} > 1$). This performance is considerably improved if the information of the surround interactions is added to the *synchrony motion maps* (case (b)), confirming its important role in motion representation.

Comparison with the analog V1-MT architecture

The similar experimental protocol allow us to have a comparison between the recognition error rates obtained in this chapter with the ones of Chapter 7. Table 8.4 shows the mean recognition error rates and standard deviations obtained for each case.

Results in Table 8.4 show that the recognition performance is better for motion maps obtained using a mean firing rate estimation, which is obtained using the analog V1-MT architecture. The results generated for mean motion maps in the spiking

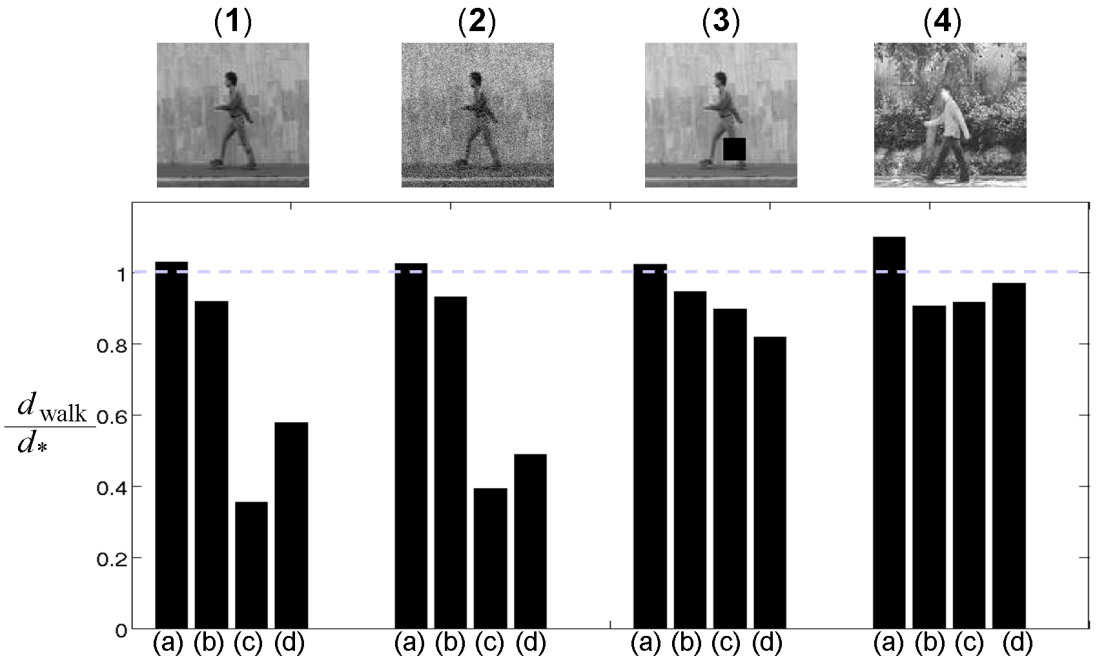


Figure 8.10: Results obtained in the robustness experiments for the four input sequences represented by the snapshots at the top of the image. From left to right: (**1**) *normalwalker*, (**2**) *noisy* sequence, (**3**) *occluded-legs* sequence and (**4**) *moving-background* sequence. For each input sequence the action recognition experiment was performed 4 times: (a) *synchrony motion maps* with MT CRF, (b) *synchrony motion maps* with MT CRF + symmetric surrounds, (c) *mean motion maps* with MT CRF and (d) *mean motion maps* with MT CRF + symmetric surrounds. The black bars indicate the ratio between the distance to *walking* class d_{walk} and the distance to the second closest class d_* (*galloping-sideways*, *bending* or *jumping-forward-in-two-legs*).

Table 8.4: Comparison of the recognition performances of the analog V1-MT architecture (Chapter 7) and the spiking V1-MT architecture (Chapter 8).

	Mean error rate $\pm$ STD
Analog V1-MT architecture	
$g^L = 0$, CRF	$5.68\%/\pm 3.8$
$g^L = 0$, CRF + isotropic surround	$5.86\%/\pm 4.2$
$g^L = 0$, CRF + isotropic/anisotropic surround	$1.72\%/\pm 2.3$
$g^L = 0.25$, CRF	$1.06\%/\pm 2.3$
$g^L = 0.25$, CRF + isotropic surround	$1.37\%/\pm 2.5$
$g^L = 0.25$, CRF + isotropic/anisotropic surround	$1.01\%/\pm 2.3$
Spiking V1-MT architecture	
Mean motion maps, CRF	$9.08\% \pm 4.40$
Mean motion maps, CRF + symmetric surrounds	$7.32\% \pm 4.62$
Synchrony motion maps, CRF	$13.89\% \pm 4.95$
Synchrony motion maps, CRF + symmetric surrounds	$7.19\% \pm 5.15$

V1-MT architecture cannot be directly associated to the results of the analog V1-MT architecture because their construction mechanism are different. Also MT neurons have in both cases different center-surround interactions.

Part III

Motion Integration

CHAPTER 9

THE ROLE OF V1 SURROUND INHIBITION IN THE SOLUTION OF MOTION INTEGRATION

Contents

OVERVIEW

From ambiguous motion signals, it is only possible to recover the component of motion perpendicular to the contour and not the real motion direction of objects, which is known as the *aperture problem*. The aperture problem can be solved integrating unambiguous motion signals, which are located at corners and end points. But, how this integration is performed by the visual system?

One possibility to study motion integration is to analyze MT neurons' response, specially their preferred direction. The preferred direction of a MT cell has been classically measured through a drifting grating, where most of the times the cell shows a clear direction selectivity. Studies, as the ones done by Pack et al. (Pack and Born (2001); Pack et al. (2004); Born et al. (2006)) show that the preferred direction can be modified depending on the input stimulus. Specifically, Pack et al. (2004) showed that the preferred direction measured using barberpoles instead of grating is biased toward perception, i.e., the side of the barberpole with the longest side. This preferred direction deviation, compared to the one measured with drifting grating, depends on the aspect ratio of the barberpole.

Here, we show that a simple mechanism, namely a delayed surround suppression in V1 neurons, can produce a significant shift in the preferred direction of MT neurons, as it is observed with barberpoles of different aspect ratios. The surround suppression acts like an end-stopping cell enhancing the responses of the V1 neurons located at the border of the barberpoles, and inhibiting the activation of the V1 cells located at the center. We also evaluated the effect of V1 surround suppression with different stimuli, such as, plaids type I, plaids type II and unikinetic plaids.

Contributions of this chapter

1. A simple mechanism to explain the shifting on the preferred-direction of MT neurons as a motion integration solution.

Keywords: aperture problem, center-surround interaction, V1 surround suppression, MT, barberpoles, plaid type II, unikinetic plaids.

Organization of this chapter

Section 9.1 describes the aperture problem and our motivation to deal with this topic. Section 9.2 describes the specific implementations for V1 and MT neurons. Section 9.3 shows the implementation details, the experimental protocol and the results obtained for barberpoles and plaids.

9.1 THE APERTURE PROBLEM

9.1.1 Definition

As previously stated in Section 4.2.1, the motion of a homogeneous contour is ambiguous. Because the receptive fields of motion sensitive neurons in the visual system are finite, each neuron is observing the world inside an "*aperture*". Within this aperture, different physical motion cannot be distinguished, as it can be seen in Figure 9.1, where different motion directions will elicit the same response in the motion sensitive neuron. Unamiguous motion can be obtained from terminators, such as end-lines or cornes. Terminators contain the real motion direction of the object. The aperture problem is solved, i.e., the motion direction of the object is correctly perceived, combining the ambiguous and unambiguous motion information, mechanism that is called *motion integration*.

This aperture problem was initially studied by Stumpf (1911) (translated version: Todorovic (1996)), and it has different variants and extensions, such as e.g., the barberpole illusion which is further discussed in the next section.

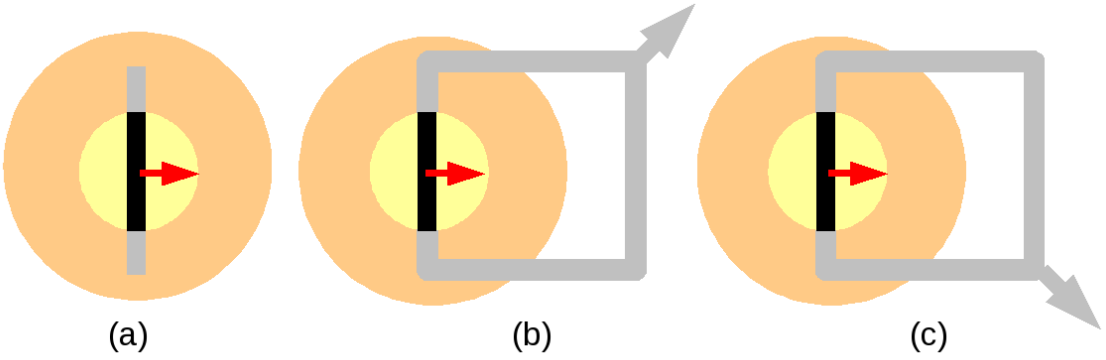


Figure 9.1: (a) The motion of a contour crossing a small yelow aperture symbolizing the receptive field of a motion sensitive neurons, typically V1. The motion direction of the contour is ambiguous and the orthogonal direction is the one eliciting the highest response. Same response is perceived for infinite motion directions and velocities, such as: (b) up-right motion, (c) down-right motion. A bigger receptive field, as the one represented in orange corresponding, e.g., to the receptive field of a MT neuron, will be able to integrate different motion cues: ambiguous 1D and unambiguous 2D. This integration mechanism solves the aperture problem finding the real motion direction of the objects.

9.1.2 The aperture problem is a motion integration problem?

One of the most popular examples about motion integration of 1D and 2D motion cues, is the barberpole illusion. The barberpole illusion is a visual illusion where a drifting grating is seen through a rectangular aperture. The motion direction perceived varies according to the relationship between the longer and shorter side (as-

pect ratio). The perception of motion is biased in the direction of the longer axis, and its deviation has been measured in psychophysics community by e.g., Wallach (1935); Wuerger et al. (1996). Figure 9.2 shows two different barberpoles with aspect ratios of 5:1 and 3:2.

Figure 9.2: Barberpole illusion represented by two different barberpoles stimuli: (a) barberpole with aspect ratio 5:1 (b) barberpole with aspect ratio 3:2. The red arrows represent the orthogonal direction of the drifting gratings.

This illusion can also be studied from a neurophysiology point of view. For example, Pack et al. (2004) measured in monkeys the preferred directions of MT neurons using barberpoles instead of classical drifting gratings. They reported that the preferred direction of MT neurons was biased towards the longest axis of the barberpole and the strength of the shifting is related to its aspect ratio, which is consistent with the perception measured by psychophysics experiments.

But, Pack et al. (2004) also showed that the preferred direction of MT neurons evolves along time, having an early response in the orthogonal direction of the drifting gratings inside the barberpole, and a late response biased the longer axis of the barberpole. Evidence of microelectrode recordings in MT of alert monkey reveal that during the first 80ms after the onset stimulus the response is strongly biased by 1D motion, i.e., the direction defined by the orthogonal direction to the contours, but lately the 2D motion direction is encoded. These experiments suggest that the aperture problem is solved within the first 100ms of the onset stimulus Pack and Born (2001). A diagram showing the evolution on time of the preferred direction of a MT neuron is shown in Figure 9.3.

The mechanisms underlying the preferred direction deviation of a MT cell are not at all defined. It looks like that the primate visual system initially considers all the information available (ambiguous and unambiguous), and that along time, it refines it in order to solve the aperture problem. This convergence in time can be associated to different and complex neural networks which convey information coming from other areas of the visual system as feedbacks (Berzhanskaya et al. (2007)) or horizontal connections. This phenomenon is also associated to the contribution of terminators or end-points in different areas of the visual field such as V2 or V1

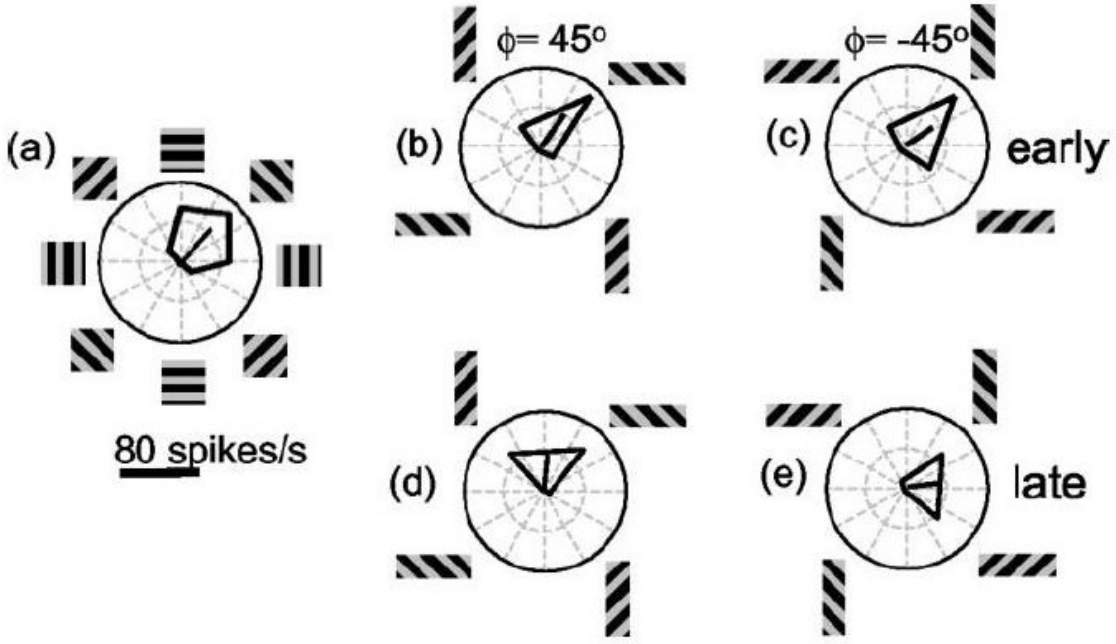


Figure 9.3: Responses of one MT neuron to barberpole stimuli. (a) Preferred direction of MT neuron measured using gratings drifting to different orientations (barberpole with aspect ratio 1:1). (b)-(c) Response of the MT neuron for the first 40msec (*early* respose). The preferred direction follows the orthogonal direction of contours and it is slightly affected by the elongation of the aperture. (d)-(e) Response of the MT neuron for longer stimulus durations (200-1000ms after the stimulus onset) (*late* responses). In this last case the preferred direction are completely affected by the elongation aperture, showing a shifting towards the longer axis of the aperture (image taken from Pack et al. (2004)).

(Berzhanskaya et al. (2007); Bayerl and Neumann (2007); Pack et al. (2003)) which should require slightly longer latencies.

Motivated by recent findings that 2D motion of terminators is faithfully encoded by V1 neurons that exhibit strong surround suppression (Pack and Born (2001); Sceniak et al. (2001); Jones et al. (2001)), we explore in this chapter the role of V1 surround suppression on the preferred direction shifting of MT cells.

9.2 Implementation of V1 and MT neurons

Starting from the definition done in Chapter 5, we present here specified models for V1 and MT neurons.

9.2.1 V1 neuron implementation

V1 neurons are here modeled as a dynamic equation where their activation is given by the value of their membrane potential $u^{V1}(t)$. The membrane potential $u_i^{V1}(t)$ of

the ith V1 neuron is then defined by

$$\frac{du_i^{V1}(t)}{dt} = -g^L u_i^{V1}(t) + k_{inh} \sum_{j \in \Phi} w_{ij} S_r(u_j^{V1}(t - \delta t)) + I_i(t), \tag{9.1}$$

where g^L and k_{inh} are the respective leak and inhibitory constants. w_{ij} defines the inhibitory connection weights representing the strength of the synapse between the ith V1 neuron and the jth V1 neuron. The delay of δt represents the delay associated to horizontal connections. $S_r(\cdot)$ is a nonlinear function used to estimate the mean firing rate of the jth V1 neuron from its membrane potential $u_j^{V1}(t)$. As nonlinearity, we use a normal rectification defined by

$$S_r(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

The external input current $I_i(t)$ contains the motion information extracted from the input stimulus, and it is defined by

$$I_i(t) = k_{exc} C_i(\mathbf{x}_i, t), \tag{9.3}$$

where k_{exc} is an excitatory amplification factor and C_i refers to the complex cell response defined in equation (5.9).

9.2.2 MT neuron implementation

The goal of this chapter is to study the role of the V1 surround suppression in the aperture problem solution. For this, MT neurons will be only considered as pooling entities with a nonlinear function at the end.

The pooling mechanism for a MT neuron is the one described in Section 5.2. Considering as output activity of the ith MT neuron (A_i^{MT}) the value of its membrane potential $u_i^{MT}(t)$, equation (5.14) can be here written as

$$u_i^{MT}(t) = \max\left(0, \sum_{j \in \Omega_i} w_{ij} S_r(u_j^{V1}(t)) - \sum_{j \in \Omega_i'} w_{ij} S_r(u_j^{V1}(t))\right), \tag{9.4}$$

where w_{ij} is the synapse weight between the jth V1 neuron and the ith MT neuron specified in equation (5.13). Ω_i and Ω_i' are the domains defined in equations (7.5) and (7.6), respectively.

$S_r(\cdot)$ is the nonlinear function defined in equation (9.2) to estimate the mean firing rate of the jth V1 neuron from its membrane potential value $u_j^{V1}(t)$.

Table 9.1: Values of the size of V1 receptive fields and V1 surrounds depending on their spatial frequency.

Spatial frequency cycles/pixel	Receptive field size	Surround size
0.04255	24	52.8
0.08510	12	26.4
0.17021	6	13.2

9.3 EXPERIMENTS

9.3.1 Implementation details

Input stimuli: We used stimuli of 200×200 pixels of: barberpoles, plaids type I, plaids type II and unikinetic plaids (see Section 9.3.3).

V1 settings: V1 layer was formed considering a homogeneous density of 0.125 cells/pixel and a radius of 90 pixels. 16 spatial orientations were implemented, as well as, 9 different spatiotemporal frequencies, giving as a results, spatiotemporal energy filters tuned at: 2, 4, 8Hz and 0.05, 0.1, 0.2 cycles/pixel. We considered $g^L = 0.05$, $k_{inh} = 0.8$ and $k_{exc} = 1$.

The surround size is defined as 2.2 times the size of the V1 receptive field, which is given by the spatial frequency of the motion energy filter associated (term $I_i(t)$ in equation (9.1). According to the spatial frequencies defined, the sizes of V1 receptive field and V1 surround are listed in Table 9.1.

MT settings: MT layers were formed by a single cell placed at the center with a receptive field covering all the input stimulus, i.e, with a radius of 100 pixels. We used 16 layers of MT neurons, each of them tuned to a different motion direction.

9.3.2 Experimental protocol

The preferred direction of a MT neuron is measured calculating its mean firing rate for input stimuli moving at different spatial directions. Considering the strength of the response for each stimulus orientation, a polar graph is obtained showing the preferred motion direction. This mechanism requires that the input stimulus must be rotated as many times as many motion orientations want to be measured. In our case, we could decide not to rotate the input stimulus but to create identical MT neurons tuned to different motion orientations (in a drifting grating sense). This procedure is equivalent to the stimulus rotation and the final preferred direction of the MT neuron will be obtained combining the responses of the different MT neuron layers.

To better visualize the effect of the surround inhibition in the output of MT neurons, the surround effect arised just after V1 membrane potential values are stabilized. The neuron response was then divided into two stages: early response (without V1 surround suppression) and *late* response (with V1 surround suppression). Figure 9.4 illustrates the *early* and *late* stages in the response of a V1 neuron.

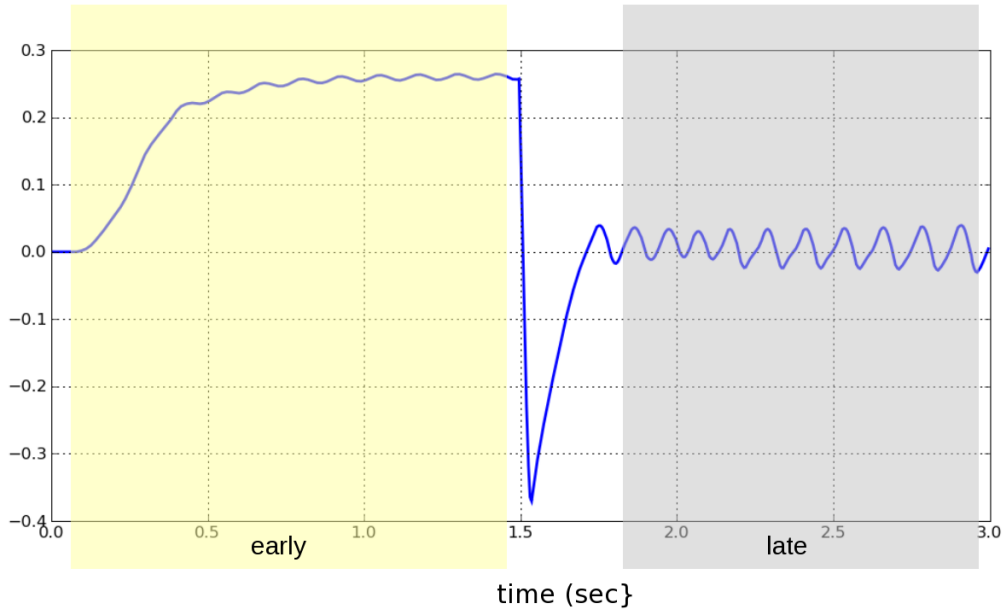


Figure 9.4: Early (yellow) and late (gray) stages of a V1 neuron response. Early stage does not consider V1 surround suppression, while in the late stage it does.

MT neuron response was also divided into two stages: *early* and *late* response. For each of them, and for a MT neuron i, the mean activation was calculated averaging in time the values of $u_i^{MT}(t)$, obtaining of this way, $\{u_i^{MT}\}_{\text{early}}$ and $\{u_i^{MT}\}_{\text{late}}$.

The *early* and *late* mean responses of the 16 MT layers were normalized dividing its activation by the sum of the mean activations of the totality of MT layers, in other words,

$$\{\tilde{u}_i^{MT}\}_{\text{early}} = \frac{\{u_i^{MT}\}_{\text{early}}}{\sum_{k=0..15}\{u_k^{MT}\}_{\text{early}}}, \qquad (9.5)$$
$$\{\tilde{u}_i^{MT}\}_{\text{late}} = \frac{\{u_i^{MT}\}_{\text{late}}}{\sum_{k=0..15}\{u_k^{MT}\}_{\text{late}}}.$$

Finally, to estimate the mean firing rate of the ith MT neuron, the values of $\{\tilde{u}_i^{MT}\}_{\text{early}}$ and $\{\tilde{u}_i^{MT}\}_{\text{late}}$ were passed through a sigmoid as the one defined in equation (7.1).

9.3.3 Results

In this section we studied motion integration dynamics of the stimuli presented in Figure 9.5, namely barberpoles and different kind of plaids. The definition of a plaid

type II will be given in the sequel.

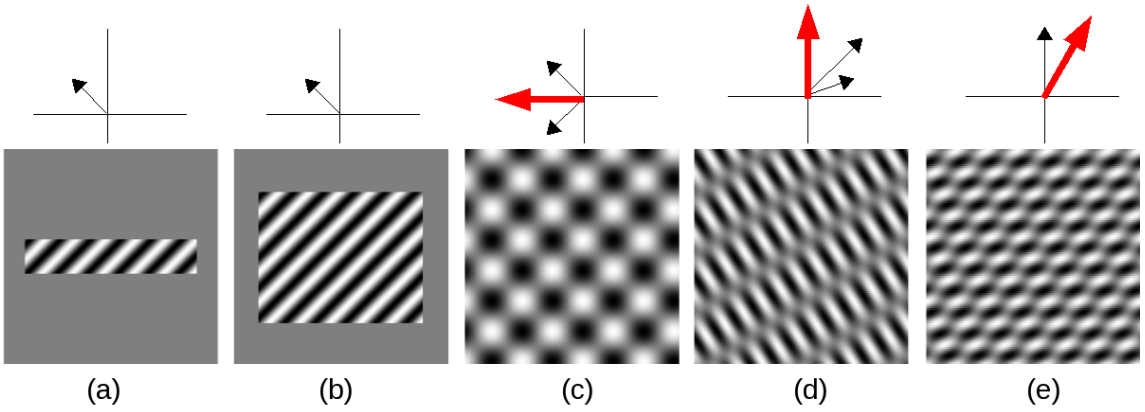


Figure 9.5: Snapshots of different stimuli used to test the V1 surround suppression effect. The respective velocity spaces are displayed at the top of each stimulus. Black arrows represents the motion direction of drifting gratings for the plaid stimuli. Red arrows symbolize the motion perceived. (a) Barberpole aspect ratio 5:1. (b) Barberpole aspect ratio 3:2 (1.5:1). (c) Plaid type I. (d) Plaid type II. (e) Unikinetic plaid.

The barbepole illusion

We asked whether the MT neurons proposed in this chapter were able to reproduce the preferred direction shifting reported by Pack et al. (2004). To do so, we tried our architecture using two different barberpoles with two different aspect ratios: 5:1 and 3:2 (1.5:1). Gratings at the center had an spatial frequency of 0.1 cycles/pixel and a temporal frequency of 4Hz. The drifting direction was 135° (see Figure 9.5).

In Figure 9.6, we observed the response of V1 neurons placed at different locations and we focused on the evolution of their response for the *early* and late stages, which means, before and after the V1 surround suppression. As it is possible to see in Figure 9.6, the response of V1 neurons are clearly affected by the surround suppression, they are biased towards the largest border of the barberpoles.

Similarly, in Figure 9.7 we observed the response of the MT cells placed at the center of the stimulus. We observed that the preferred direction, after the V1 surround suppression, is clearly shifted towards the longer axis of the barberpole. This effect is stronger for higher aspect ratios, i.e., the preferred direction shifting obtained for the barberpole with aspect ratio 5:1 is bigger than the preferred direction shifting obtained for the barberpole of aspect ratio of 3:2.

Plaids: type I, type II and unikinetic

We also tested the V1 surround suppression with other stimuli, such as plaids type I, plaids type II and unikinetic plaids. Stimuli are shown in Figure 9.5 (c-e) and the results are shown in Figure 9.8.

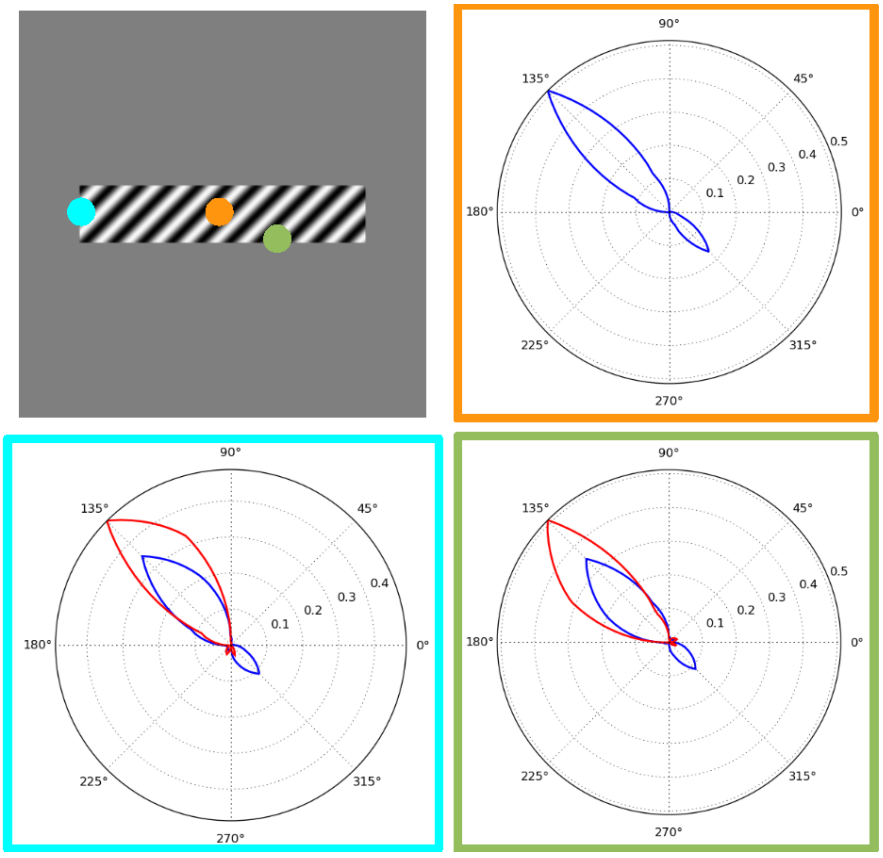


Figure 9.6: Output of V1 neurons for the barberpole illusion. The barberpole has an aspect ratio of 5:1 and the grating drifts to 135°. The figure shows the response of V1 cells located at three different places: center (orange), vertical border (cyan) and horizontal border (green). The blue line represents the output of V1 neurons for the *early* response. Red lines represent the output of V1 neurons for the *late* response, which is clearly shifted to the largest border direction. In the case of the V1 cells placed at the center, the *late* response does not exist because their activation are completely inhibited by the surround suppression.

The plaid type I is formed by two gratings drifting to 135° and 225°, respectively (see Figure 9.5 (c)). Both gratings have identical spatial and temporal frequencies: 0.1 cycles/pixel and 4Hz. In plaid type I stimulus, the perceived direction, IOC and VA coincides, and the surround suppression has no special effect in the MT preferred direction, as it is shown in Figure 9.8 (a-c).

In the case of plaids type II, we tested a plaid formed by one grating drifting to 135°, 6Hz and 0.1 cycles/pixel and a second grating drifting to 160°, 3Hz and 0.1 cycles/pixels (see Figure 9.5 (d)). In this case, the VA solution significantly differs from IOC solution, being this last one closer to the perceived direction. In this case the surround inhibition did not showed special effect in the MT preferred direction (see Figure 9.8 (d-f)).

For the unikinetic plaids, we used one static grating spatially oriented at 135 °and a drifting grating drifting in the 90° direction with a temporal frequency of 3Hz (see Figure 9.5 (e)). Both gratings have a spatial frequency of 0.1 cycles/pixel. The uniki-

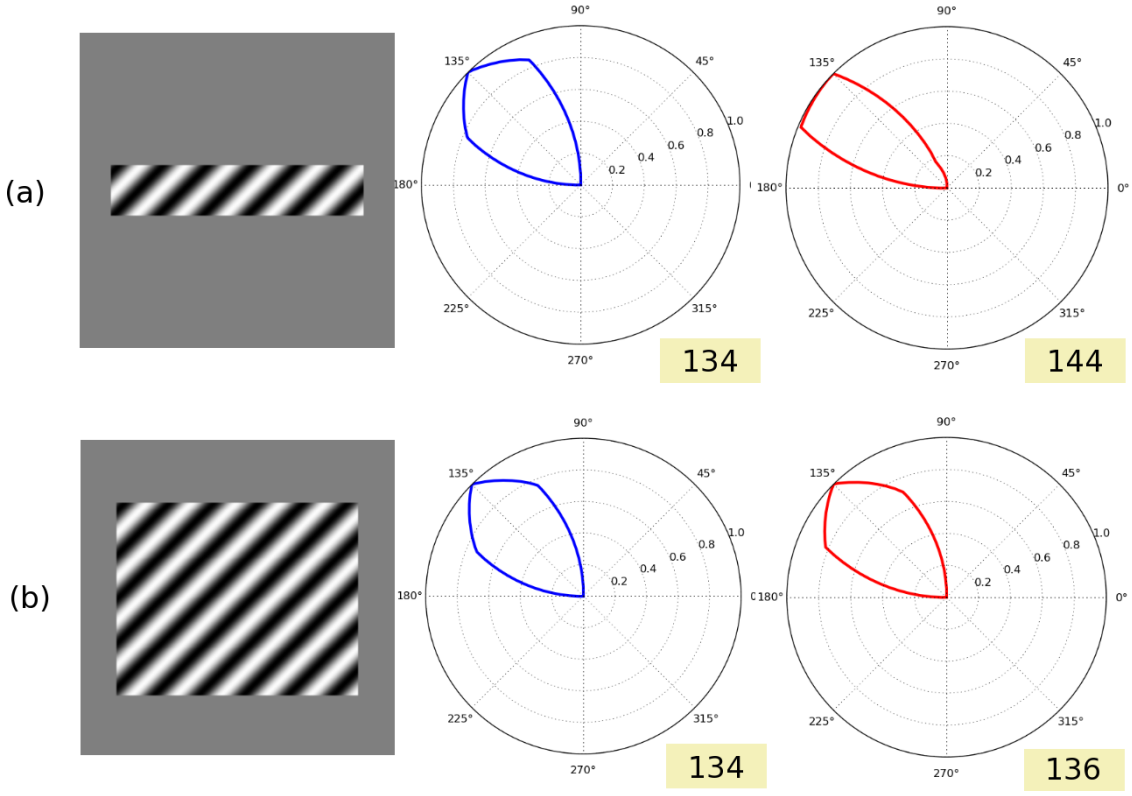


Figure 9.7: Output of MT neurons for the barberpole illusion. The polar graphs show the response of the 16 MT neurons implemented. The first column shows a snapshot of the input stimulus. Second column shows the polar graph for the *early* response (blue line). Third column shows the polar graph for the *late* response (red line). (a) Barberpole with an aspect ratio of 5:1 drifting in 135°direction. (b) Barberpole with an aspect ratio of 3:2 drifting in 135°direction. For the *early* and *late* stages, and for each barberpole, the value of the MT preferred direction (in degrees) is displayed in the yellow box at the bottom of each polar graph. The preferred direction shifting of MT neurons significantly changes for the barberpole of aspect ratio 5:1 towards perception. The results obtained for the barberpole of aspect ratio 3:2 show a slight preferred direction shifting.

netic plaids are treated in a different manner. How only one grating component drifts, the IOC and VA mechanisms cannot be applied. In our simulations we perceive a shifting in the MT preferred direction of around 10° towards the perceived direction (see Figure 9.8 (g-i)).

The deviation observed in the preferred direction of MT neurons in our model (about $10°$) is smaller than found by Pack et al. (2004) using barber-poles (about $25 - 30°$ in about 150ms). This indicates that the weight given to 2D motion information is still not sufficient to indicate the true global motion in our model

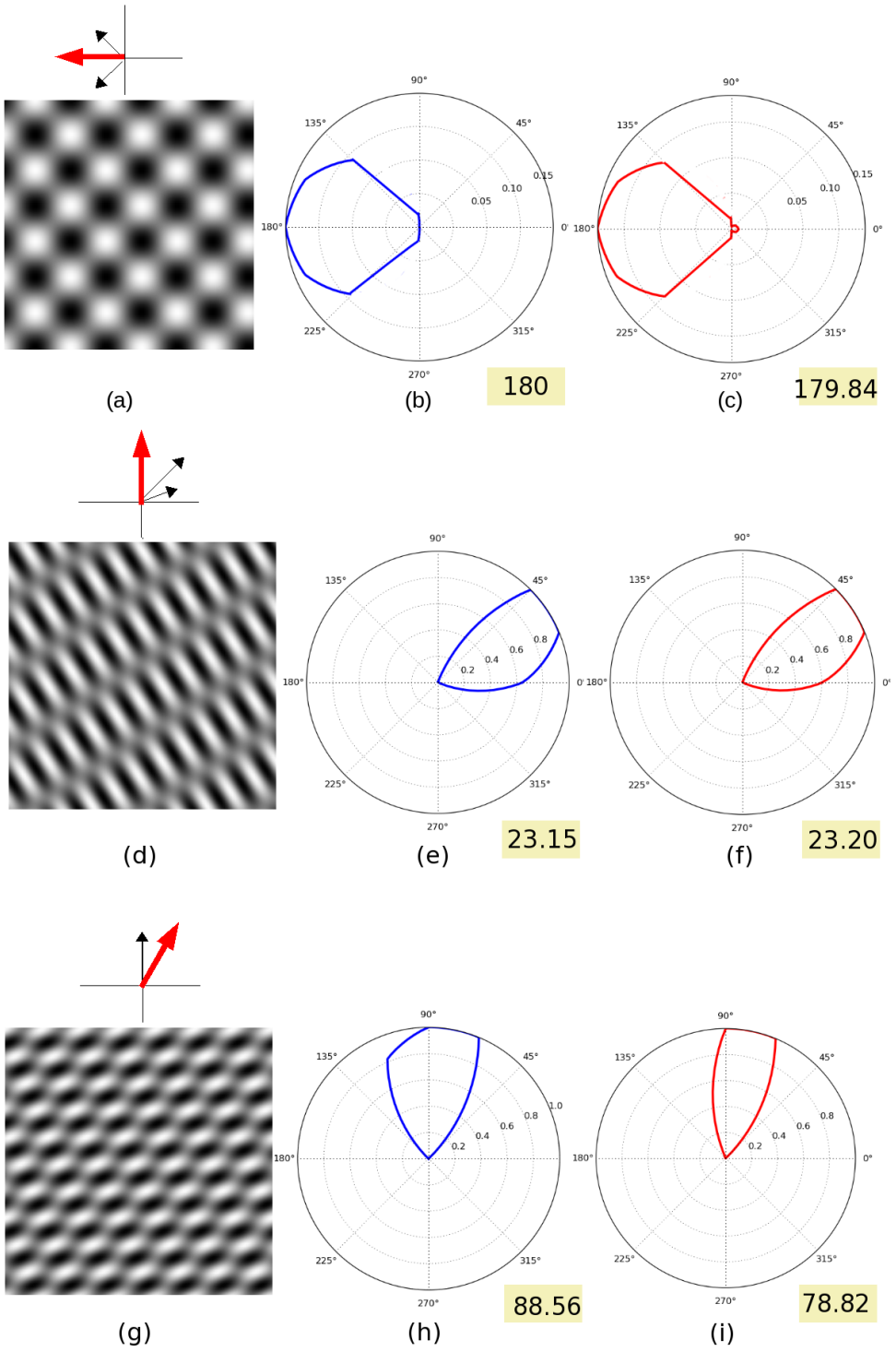


Figure 9.8: Effect of V1 surround suppression in the preferred direction of a MT neuron. For each stimulus and for the *early* and *late* stages, the value of the preferred direction of MT neurons (in degrees) is displayed in the yellow boxes at the bottom of each polar graph. The effect is measured on three kinds of plaids presented in the first column: (a) plaid type I, (d) plaid type II, (g) unikinetic plaid. The second column shows the *early* response while the last column shows the *late* response. The shifting in the preferred direction of MT cells is only clear on the unikinetic plaid.

Part IV

Conclusion

CHAPTER 10

CONCLUSION

"Great is the art of beginning, but greater is the art of ending"
– Henry Wadsworth Longfellow

10.1 SUMMARY

In this thesis we studied the motion perception in mammals and how bio-inspired systems can be applied to real applications on real image sequences. Using properties of directionally-selective neurons in V1 and MT macaque brain areas, we proposed a feedforward V1-MT core architecture for motion processing. This architecture is formed by two layers of motion sensitive neurons, V1 and MT layer. We proposed two implementations for V1 and MT neurons: an analog and a spiking implementation.

10.1.1 Detecting motion

For both implementations, we defined energy motion detectors in order to extract the motion information from input image sequences. The energy motion detectors were implemented following the physiologically plausible version proposed by Adelson and Bergen (1985). This implementation shares properties with V1 directionally-selective neurons and is characterized by the fact that the tuning frequencies cannot be found analytically. We also proposed a numerical analysis of the frequency response of these filters, showing a table with the needed parameters to create filters with suitable frequencies tuning.

10.1.2 V1-MT analog architecture

From the output of the energy motion detectors we proposed two implementations for the analog V1 neurons.

- In Chapter 7 we considered directly the output of the motion detector units as the membrane potential of V1 neurons, which after a nonlinearity, it can be

interpreted as an estimation of its mean firing rate.

- In Chapter 9, a V1 neuron is implemented as a dynamical differential equation where the value of its membrane potential evolves along time. In this case, the output of motion energy filters is an external input current of the neuron, which also has inhibitory neighboring connections with the neurons conforming its suppressive surround. Similarly to the first implementation, the value of the membrane potential is passed through a nonlinear function in order to estimate the mean firing rate.

In both cases, the estimated mean firing rates of V1 neurons feed the next MT analog layer. Each MT neuron conforming the MT layer, has a receptive field around 10 times bigger than the size of V1 receptive fields. We also implemented two types of analog MT neurons, depending on the applications we considered:

- **The role of different MT center-surround configurations:** This goal was assessed in Chapter 7, where we studied the effect of different center-surround configurations in the performance human action recognition. MT neurons were modeled as a conductance-based neurons where the value of their membrane potential evolves along time according to input conductances. The value of the input conductances depends on the mean firing rates of V1 neurons. Each MT neuron has an input excitatory conductance which is obtained considering the mean firing rates of V1 neurons inside its classical receptive field, and with an absolute difference of angle between the preferred directions of V1 and MT neurons less than $\pi/4$. Then, the inhibitory conductance was obtained considering the V1 neurons whose receptive fields are centered at the inhibitory region of MT neurons. Analogously to V1 neurons, the membrane potential value of MT neuron was passed through a nonlinear function in order to estimate the value of its mean firing rate.

- **The role of V1 surround-suppression in the solution of the aperture problem:** This goal was assessed in Chapter 9, where we studied the role of V1 surround-suppression in the motion integration problem. V1 modeling was important as we previously described, and for MT we only implemented a pooling entity followed by a normalization and a subsequent nonlinear function. The pooling algorithm was the same described for the excitatory conductance of the previous paragraph.

10.1.3 V1-MT spiking architecture

In the analog implementation we assumed that the mean firing rate was a sufficient representation of the activation of V1 and MT neurons. But, real neurons communicate through spikes, and spike trains contain much more information than just the mean firing rate. For example, this was exemplified by the work of Thorpe et al. (see

e.g., Thorpe et al. (2001); VanRullen and Thorpe (2002)) and Gollisch and Meister (2008), who showed that before the mean firing rate is available, the visual system already received important information about the input stimulus. Following this motivation, we proposed in Chapter 8 a fully spiking V1 and MT layers. In this spiking implementation, V1 neurons are modeled as integrate-and-fire neurons where the output of motion energy filters feed the V1 neurons as an external input current. The output of this spiking V1 neurons, *spike trains*, are directly transmitted to the next spiking MT layer. The subsequent spiking MT neurons are modeled as conductance-driven integrate-and-fire neurons, where the spikes generated by afferent V1 neurons form the excitatory input conductance. Similarly to the MT analog neurons, we also implemented different inhibitory center-surround interactions. So, inhibitory conductance was formed by the spikes generated by V1 neurons falling into the inhibitory MT surround area.

A major difference with the analog implementation of V1 and MT neurons, is that no nonlinear function was added at the output of each neuron. This is mainly because the spike generating process is already nonlinear and the mean firing rate can be directly estimated counting the number of spikes emitted inside a time window.

10.1.4 Recognizing human actions

One of the goals to propose bio-inspired feedforward V1-MT models was to apply them to a real application: human action recognition in real scenes.

To do so, from the output of our system, i.e., from the output of MT neurons, we defined features vectors representing the input sequences, and with those feature vectors we performed recognition.

With the analog V1-MT architecture, we proposed a *mean motion map* as a representation of the motion information of the input stimulus. This mean motion map is defined as a vector with a length equals to the total number of MT neuron. Each position of the vector contains the estimated mean firing rate of each MT neuron inside a temporal window. Using these *mean motion maps* we also evaluated the effect of different MT surround geometries in the human action recognition performance.

With the spiking V1-MT architecture, we proposed two different motion maps, the two of them representing different characterizations of the spike trains: *mean motion map* and *synchrony motion map*. Similarly to the *mean motion map* defined for the analog V1-MT architecture, the *mean motion map* for the spiking architecture is formed calculating the mean firing rate of each MT neuron inside a time window. Then, the *synchrony motion map* does not code the mean firing rate, but the synchrony between them. The *synchrony motion map* is an array of matrices where each matrix represents the synchrony between neurons of the same MT layer. We had many layers as many direction-selectivity and center-surround interactions defined. The value of the synchrony is calculated using the ISI representation and the metric defined by Kreuz et al. (2007).

10.1.5 Solving the aperture problem

The other application treated in this thesis was the study of the effect of the V1 surround suppression in the solution of the aperture problem. In other words, how the 2D information extracted by the V1 surround suppression mechanism can be integrated by a MT neuron to solve the aperture problem.

The surround suppression, modeled by an isotropic difference of Gaussians (DoG), acts like an end-stopping cell (Sceniak et al. (2001); Jones et al. (2001)) enhancing the responses of the V1 cells located, e.g., at the border of a barberpole stimulus, and inhibiting the activation of the V1 cells located at the center.

To do so, we tested different psychophysical stimuli such as barberpoles, plaid type I, plaid type II and unikinetic plaid. With those stimuli, we observed the effect of a delayed V1 surround suppression obtaining the MT preferred direction.

10.2 DISCUSSION

In this section, our goal is to make a critical analysis of the concepts and results presented therein, but also to discuss the potential avenues of this work for future research efforts.

10.2.1 V1-MT modeling

In V1 motion is processed by directionally-selective neurons, which can be found in V1 simple and complex cells. The directionally-selective property is specially enhanced in MT neurons, which also tend to have higher preferred speeds compared to V1. In V1 complex cells is possible to find neurons exhibiting spatiotemporal response which does not depend on the spatial frequency of the input stimulus (Priebe et al. (2006)). This type of neurons has been also found in MT (Priebe et al. (2003); Perrone and Thiele (2001)). Neurons with this property, also known as *speed-tuned* neurons have been the target of different V1-MT models proposed in the literature (see e.g., Grzywacz and Yuille (1990); Simoncelli and Heeger (1998); Perrone (2004).

In our case we did not model *speed-tuned* neurons. Our V1 neurons, which are based in the energy motion detectors proposed by Adelson and Bergen (1985), are tuned for a certain speed but in a very limited spatiotemporal bandwidth. MT neurons in our case do not pool V1 neurons in order to obtain a plane tuned for a certain speed, as done by Simoncelli and Heeger (1998) and Grzywacz and Yuille (1990), who proposed different algorithms to find the same velocity plane. MT neurons pool V1 neurons with the same motion direction selectivity, grouping of this way, the responses of all the spatiotemporal frequencies of the space. This simplified pooling mechanism was basically inspired by the application of human action recognition, where studies such as Casile and Giese (2005), have shown that our ability to recog-

nize actions require a minimal speed information compared to motion direction and a coarse spatial location of the motion information.

▼

A natural extension of our work would be to consider the implementation of ***speed-tuned neurons****. This type of neurons will be necessary to have a system correctly handling different phenomenons in V1 and MT neurons, such e.g., plaids type II and pattern-component neurons.*

V1 motion detection

For early local motion detection, Simoncelli and Heeger (1998) proposed local units modeled through spatiotemporal energy filters. However, those filters have a temporal profile that is non-causal and inconsistent with V1 cell physiology. Our approach, on the other hand, uses temporal profiles consistent with V1 cell physiology. These biologically plausible temporal profiles bring out not trivial calculation for the tuning of the spatial-temporal frequency orientation. As a consequence, motion orientation tuning must be computed using numerical approximations.

MT pooling mechanism

In this thesis, we first presented MT neuron as a generic pooling entity which renders the output information of the V1 complex cells. We show that the pooling mechanism gives as result MT neurons with a high direction-selectivity, which is an important property present in real MT neurons (Albright (1984); Snowden et al. (1991); Lagae et al. (1993); Churchland et al. (2005)), but we do not obtain *speed-tuned* neurons, as we previously discussed. The dynamic activity of a MT neuron will depend on the V1 neuron models implemented. For instance, in Chapter 9 MT neurons only pools the mean firing rate of V1 neurons, in Chapter 8 MT neurons integrates the spikes generated by V1 neurons, while in Chapter 7 MT neurons only integrate the mean firing rate of V1 neurons which are analog values.

Center-surround interactions

The majority of V1 simple and complex cells have center-surround interactions, which can be facilitatory or suppressive depending on the properties of the input stimulus, i.e., contrast, spatiotemporal frequency, stimulus type (bars, random-dots, plaid, barberpoles, etc.). Analogously, MT cells also exhibit center-surround interactions which also depend on the input stimulus. Suppressive-surround mechanism is supposed to be involved in the solution of the aperture problem modulating the preferred direction of MT neurons. We presented different center-surround interactions in MT neurons which were implemented in Chapters 7 and 8. We showed that this interaction, which is normally suppressive, codes important motion singularities that will be crucial for motion categorization in a real vision application, as e.g. for the human

action recognition application. We also implemented center-surround interaction in V1 neurons (Chapter 9), showing that V1 surround suppression can extract 2D information and contribute, of this way, with the solution of the aperture problem. We did not try different V1 surround configurations.

Our model implements different MT non classical receptive fields by having different classes of center-surround interactions (e.g. Xiao et al. (1995), Born (2000)). The role of different MT receptive field shapes in the action recognition task has not been evaluated before. Here we present some results in the action recognition performance using different structures and geometries of receptive fields as observed in monkey area MT (Born (2000); Xiao et al. (1995, 1997b)), showing their crucial role in our motion representation.

Feedbacks?

The quote at the beginning of Chapter 5[1] really inspired our work and it explains the reason why no feedback mechanisms were included in this model.

▼

*Motion analysis can be performed without **feedbacks**, but of course, feedbacks could give to our model additional capabilities and robustness. Considering feedbacks*[2] *was outside of the scope of this thesis, and this is also a natural perspective. For instance, our model is not able to deal with crowding, or it has difficulties to treat occlusions or complex backgrounds, tasks that can also been implemented using attentional mechanisms. For the resolution of the aperture problem feedbacks also play an important role, e.g., diffusing the non-ambiguous information seen by upper layers to neurons in a lower layer to enhance the real motion direction of the objects and to solve the aperture problem.*

Towards a modelization of the visual system?

In this thesis, we focused on the modelization of a small piece of the visual system, specially if we look back at the Felleman & Van Essen diagram of the visual system shown in the introduction. So we can wonder, how this contribution could be extended or included in a larger model of visual system?

▼

For example, our system deals with motion analysis considering that input images arrive directly to V1, where a first filtering stage is done. But this simplification lacks a non neglected part of the visual system, which is all the way from the retina to V1. So, the first idea that comes in mind would be to connect our system to the output of

[1] *"We cannot think about what a feedback interaction could do if we do not first explore the limitations of a feedforward model"* – Simon Thorpe (GDR-vision meeting 2008)

[2] Note that the inclusion of feedback connections is not an original idea, since it has been already implemented in several motion models processing (see Chapter 4). However, it could be interesting to implement feedbacks in our model to investigate its different roles in motion processing and integration.

a retina model (see e.g., Hérault and Durette (2007); Wohrer and Kornprobst (2009)). But then, what will be the gain with this new architecture? Maybe, a simplified retina performing operations, such as edge detection and contrast-gain normalization, would be sufficient.

But more generally, a challenging perspective is to consider our contribution as a part of a global model of the visual system. The integration of different elements of the visual system requires non-trivial efforts to understand and to implement the connectivity between different layers. A big effort in this direction has been done by the BlueBrainProject[3]*, who attempts to model neocortical columns that can be used to simulate any brain area.*

To deal with this long term perspective, many questions will require some attention: What kind of mathematical framework is the most suitable? How to deal with the fusion of different scales? What are the computational challenges for the implementation?

10.2.2 Human action recognition: result analysis

Analogue or spike architecture?

With the analogue architecture, we showed that the inclusion of different center-surround interactions in MT neurons can significantly improve the recognition performances. We think that different center-surround interactions in MT neurons extract singular motion patters that act like key information for motion categorization task.

With the spiking architecture, recognition results obtained using *synchrony motion maps* are slightly inferior than the ones obtained using *mean motion maps*, specially if we only consider the activation of MT CRFs. This difference is enhanced in the robustness experiments. As an explanation, we think that because the synchrony analysis largely forgets about the rate, it lacks a fundamental information about network activity. Nevertheless, by considering synchronies only, satisfying recognition performance can be achieved. Also, note that the use of the synchrony to encode the input motion information improves the inter-class separability obtaining a better class clustering (see Figure 8.9 and Table 8.3). These results are consistent with neuroscience findings about the complementarity of rate and synchrony codes: There are evidence from motor and visual cortex that both, rate and synchrony code, are conjointly used to extract complementary information, (Maldonado et al. (2008); Grammont and Riehle (2003); Riehle et al. (1997)). As a future work, we plan to combine these two motion maps in order to have a better representation of the input motion information.

Although the *mean motion maps* of the analogue and spiking architecture have the same philosophy of construction, there are differences in the recognition error

[3] `http://bluebrain.epfl.ch`

rates obtained for both architectures. We think that this difference is due mainly because the parameters of the model are not the same and the spike generation mechanisms in both representations are not equivalent.

In general, the results obtained with the spiking architecture are not so good as the ones obtained with the analog architecture. This does not mean that the use of spikes does not bring anything new, it means that either we are not interpreting the neural code in a proper manner or the motion maps proposed as representations of the input motion information are not really representatives.

▼

Regarding to spiking architecture, the main perspective will be the study of ***higher order statistics of the spike trains*** *generated by MT neurons. It is likely that this study will give us new insights about how to analyze the MT output, i.e., the motion content in videos.*

More validations?

Of course, more validation would be needed! We tested the model with Weizmann database. The good recognition performance obtained with our model, both with analog and spiking architectures, reinforces our hypothesis about the representation of our motion maps.

Weizmann database was also used by, e.g., Blank et al. (2005) and Jhuang et al. (2007) to validate their model. However, test conditions and experimental protocol are not the same than the ones considered in our experiments, and therefore recognition performances cannot be easily compared.

Based on our results, we do not claim that our system will work in any condition. But that concern is in fact general as remarked by Pinto et al. (2008): It is an overclaim to declare that the whole action recognition problem is solved only based on some results obtained with a given database, real conditions as complex background, rotations, occlusions, distractors are generally not included. So, more validation with other databases, such as KTH[4] database, would be needed.

▼

Beside doing more validations, we also think about two additional perspectives. The first perspective is to investigate how local ***form information*** *can be dynamically merged and integrated with the motion pathway to improve the representation of motion maps, specially in the case of complex backgrounds where motion integration could play an important role. The second perspective is to investigate the* ***role of classifiers*** *in the results. The representation of our motion maps clearly affects the recognition performance. But, the classifier should be also a critical element in the system. As a perspective, it could be interesting to make a benchmark of different classifiers in order to evaluate their impact in the recognition performance.*

[4] http://www.nada.kth.se/cvap/actions/

Also, we thought about additional experiments that could be done. For example, we wonder about the existence of ***keyframes.*** *It has been shown that not all the frames of an action have the same relevance for recognition. If only a few keyframes are shown, we can successfully recognize actions. It could be interesting to implement, according to the "complexity" of motion information of each frame, an automatic extraction of keyframes in order to only compare the most relevant motion information of each sequence. Another set of experiments is the* ***early recognition****. Following the spirit of ultra fast classification based mainly in Thorpe et al. (2001) and Thorpe and Fabre-Thorpe (2001), we could study the evolution of the recognition performance along time. In other words, how much time is needed to have successful recognition performance (the size of the temporal window, Δt, in the construction of motion maps)?. This study could be seen as an application of rank-order-coding for videos. Changing the value of Δt could be also interpreted as the "memory" of the system, i.e., how much information from the past is needed to have a good representation of a certain action?*

How to compare our model with existing models?

Earlier models have suggested that biological motion perception depends on strong interactions between motion and form pathways (see Blake and Shiffrar (2007) for a review). In the model proposed by Giese and Poggio (2003), both form and motion pathways learn sequences or "snapshots" of human shapes and optic flow patterns, respectively. Several models have been proposed to dynamically constrain such motion model using local information about shapes, features and contours (e.g., Bayerl and Neumann (2007); Tlapale et al. (2008)). Since configural information are important for biological motion recognition (e.g., Hiris et al. (2005)).

Specifically, Giese and Poggio (2003) proposed a neurophysiological model for the visual information processing in the dorsal (*motion*) and ventral (*form*) pathways. The model is validated in the action recognition task using as input stimulus stick figures constructed from real sequences. Assuming no interaction between the two pathways, they found that both motion and form pathways are capable to perform action recognition. Moreover, their model exhibit several interesting properties for biological pattern motion recognition such as spatial and temporal scale invariance, robustness to noise added to point-light motion stimuli and so on. One of the main difference with our approach is the fact that new parameters need to be fitted if a new action must be considered. In our case, no parameters must be adjusted and only the new motion maps must be inserted into the training set.

More recent work from Jhuang et al. (2007) implemented this invariance for spatial and temporal scales (i.e. stimulus size and execution time, respectively). Their approach uses a bio-inspired model for the action recognition task based in Giese and Poggio (2003) and Serre et al. (2005). The invariance to spatial and temporal scale is achieved considering as many motion detector layers as the number of

spatial and temporal scales to be detected followed by a *max* operator. This can be easily implemented in our approach adding more layers with different spatial and temporal scales and therefore apply the *max* operator between the different layers coding the same motion direction.

Unlike optical-flow based models, where a single velocity is assigned to each point, our model reproduces to some extent the richness of center-surround interactions specifically varying the geometries of the surrounds (see Figure 5.14) (Born (2000); Xiao et al. (1995, 1997b)). The different surround geometries give different kinds of motion contrasts for several orientations at every point. Interestingly, we showed that taking into account this diversity of MT cells improves the recognition performance. Our interpretation is that cells with inhibitory surrounds bring information related to velocity opponency or singularities in the velocity field of the input stimulus.

Contrarily to the bio-inspired model of Giese and Poggio (2003), our model relies on a general purpose motion processing based upon the known properties of the two-stages biological motion pathway where V1 and MT neurons implement detection and integration stage, respectively. The architecture is rooted on the linear-nonlinear ("L-N") model, of a kind that is increasingly used in sensory neuroscience (see Simoncelli and Heeger (1998), Rust et al. (2006) for instance). Recent version of this L-N models propose that complex motion analysis can be done through a cascade of L-N steps, followed by a Poisson spiking generation process Rust et al. (2006). Our generic motion model departs from this cascaded L-N model in several important way.

Regarding quantitative comparison, we followed the experimental protocol presented by Jhuang et al. (2007) in order to compare the results obtained. For our both architectures, the results obtained using *mean motion maps* and *synchrony motion maps* revealed a high variability depending on the sequences taken to build the *training set*. Due to the high variability found in our results, the direct comparison with Jhuang et al. (2007) is not evident.and their recognition percentages no representatives.

10.2.3 V1 surround suppression: result analysis

As we previously stated in Chapter 9, we obtained smaller shifting in the preferred directions of MT neurons than the ones reported by Pack et al. (2004), so we need additional mechanisms to really solve the aperture problem. To reach a satisfying solution of the aperture problem, several solutions will be explored

- Changes in preferred direction in MT neurons depend on several parameters of the non-linear stage at the end of V1 processing. This is consistent with the finding of Rust et al. (2006) that pattern-selectivity in MT neurons depends critically on the nonlinear processing (i.e., divisive normalization) at both V1

and MT level. Nonlinearities seem to play a fundamental role which needs to be clarified.

- Different types of center-surround interactions working at different spatial scales and with different relative orientations might be an alternative to extract 2D features motion.
- A better diffusion process of 2D information can be achieved through anisotropic interactions between different locations (see, e.g., Tlapale et al. (2008)).
- The 2D motion information extracted by the V1 surround suppression mechanism. In the case of plaids type II or unikinetic plaids, the motion perceived could be associated to a motion of a pattern that has a lower frequency compared to the drifting gratings used to create the stimuli (see Wilson et al. (1992) and Non-Fourier motion). The spatiotemporal frequency tuning of V1 neurons used to extract the 2D information is crucial to detect this pattern, and to have of this way a high response to the 2D motion cues.

We also explored the effect of V1 surround suppression in different stimuli such as plaids type I, plaids type II and unikinetic plaids. In the only plaid where a shifting in the preferred direction of MT neuron was perceived is for the unikinetic plaid, where a shifting of around 10° was obtained. We expected also to have a shifting for plaids type II, but we observed in our experiments that the V1 surround suppression did not affect the preferred direction of the MT neuron. We believe that this effect is related to the right spatiotemporal frequency that must be "seen" by our V1 motion detectors, where the spatiotemporal frequency tuning of V1 neurons becomes essential to correctly extract the 2D motion information.

▼

The model for V1 neuron defined in Chapter 9 can be formalized as neural fields where several theoretical and experimental results are established (see, e.g., Faugeras et al. (2009); Giese (1998)). Indeed, equation (9.1) *is equivalent to the voltage based model framework, where the population of neurons in our case is only the combination of four simple cells (see equation* (5.9)*). Neural field implementation requires a deeper mathematical analysis for dynamical systems, such as stationary solutions, stability, bifurcation diagrams, etc. For example, it would be very interesting to study how the stability of the system, or the number of solutions (multistability) varies, e.g., with different center-surround configurations.*

10.2.4 Software contribution

A substantial effort was done to implement code dealing with networks of neurons, spiking neurons, parallel processing, motion-energy filtering, etc.

In particular, an effort was made to implement properly the spatiotemporal filtering needed for the motion energy computation, specially because this stage represents

the most demanding calculation. This part of the work will be available soon as an open source C/C++ library.

The spiking neuron implementation was done thanks to the MVAspike library developed by Rochel (2004). Which allowed us to easily create layers of spiking neurons and connections between them.

CHAPTER 11

CONCLUSION (FRANÇAIS)

"Great is the art of beginning, but greater is the art of ending"
– Henry Wadsworth Longfellow

11.1 RÉSUMÉ

Dans cette thèse nous avons étudié la perception du mouvement chez le mammifère. Nous montrons aussi comment un système bio-inspiré peut être utilisée dans le cadre d'une application qui travaille avec des séquences d'images réelles. À partir des propriétés de selectivité à l'orientation des neurones de V1 et MT, nous avons proposé une architecture sequentielle générale, modélisant les aires corticales V1 et MT qui a comme obtectif le traitement du mouvement. Cette architecture est formée par deux couches de neurones sensibles à la direction du mouvement: la couche V1 et la couche MT. Nous avons proposé deux implémentations des neurones de V1 et MT: analogique et évènementielle.

11.1.1 Detection du mouvement

Pour ces deux implémentations, nous avons défini des détecteurs de mouvement basés sur l'energie pour obtenir l'information du mouvement à partir de la séquence d'images d'entrée. Les détecteurs de mouvement basés sur l'énergie ont été implementés en se basant sur la version physiologique de Adelson and Bergen (1985). Cette implémentation a des propriétés en commun avec les neurones de V1 sensibles à la direction, et est characterisée par le fait que le réglage des fréquences spatio-temporelles ne peut pas être fait de façon analytique. Nous avons aussi proposé une analyse numérique pour trouver la réponse en fréquence de ces filtres: Nous donnons un tableau avec les paramètres requis pour créer des filtres avec une sélectivité souhaitée.

11.1.2 L'architecture analogique de V1 et MT

À partir de la sortie des détecteurs de mouvement basés sur l'énergie, nous avons proposé deux implémentations pour les neurones de V1:

- Dans le Chapitre 7, nous avons directement pris la sortie des détecteurs de mouvement à savoir le voltage de membrane d'un neurone de V1 lequel, après une non-linéarité, peut être interprété comme une estimation du taux de décharge moyen.

- Dans le Chapitre 9, chaque neurone de V1 est implementé par une équation différentielle dynamique, où la valeur du potentiel de membrane évolue au cours de temps. Dans ce cas, le courant externe du neurone est la sortie des filtres de mouvement. Le neurone a aussi des connexions latérales inhibitrices dans un voisinage. De la même façon que l'implémentation du Chapitre 7, la valeur du potentiel de membrane est passée au travers d'une fonction non-linéaire pour obtenir une estimation du taux de décharge moyen.

Dans ces deux cas, l'estimation du taux de décharge moyen des neurones de V1 est l'entrée de la couche analogique suivante, MT. Chaque neurone de MT conformant sa couche, a un champ récepteur d'une taille d'environ 10 fois le rayon des champs récepteurs des neurones de V1. Nous avons aussi implémenté deux types de neurones analogiques pour MT, en fonction de l'application considerée:

- **Le rôle des différentes configurations de centre-périphérie pour les neurones de MT:** Cet objectif est abordé dans le Chapitre 7, où nous avons étudié l'effet de différentes configurations de centre-périphérie sur les performance de reconnaissance d'action. Les neurones de MT ont été modélisés comme des neurones à conductance, où la valeur du potentiel de membrane évolue au cours du temps selon ses conductances d'entrée. Les valeurs des conductances d'entrée dépendent du taux de décharge moyen des neurons de V1. Chaque neurone de MT a comme conductance d'entrée excitatrice les estimations des taux de décharge moyens pour les neurones de V1 dans son champ récepteur classique, et avec une différence absolue de l'angle entre les directions préférées des neurones de V1 et MT inférieur à $\pi/4$. Par contre, la conductance inhibitrice est obtenue à partir des neurones de V1 localisés dans la région inhibitrice du champ récepteur du neurone de MT. De la même manière que les neurones de V1, la valeur du potentiel de membrane d'un neurone de MT est passée par une fonction non-linéaire afin d'obtenir une estimation du taux de décharge moyen.

- **Le rôle de la suppression périphérique des neurones de V1 pour résoudre le problème d'ouverture:** Cet objectif est abordé dans le Chapitre 9, où nous avons étudié le rôle de la suppression périphérique des neurones

de V1 sur le problème d'intégration du mouvement. La modélisation de V1 a déjà été décrite, et pour MT nous avons simplement implémenté des entités de groupage, suivi par une normalisation et une fonction nonlinéaire. L'algorithme de groupage a été déjà décrit dans le paragraphe précédent.

11.1.3 L'architecture évènementielle de V1 et MT

Dans l'implémentation analogique nous avons supposé que l'activation des neurones de V1 et MT peut être réprésentée par le taux de décharge moyen. Mais, dans la réalité les neurones se communiquent par *spikes*, et les trains de *spikes* contiennent plus d'information que le taux de décharge moyen. Cette idée a par example été illustrée par Thorpe et al. (voir Thorpe et al. (2001); VanRullen and Thorpe (2002)) et Gollisch and Meister (2008), qui ont montré qu'avant que le taux de décharge ne soit disponible, le système visuel a déjà reçu des informations importantes sur le stimulus d'entrée. Motivés par ce fait, nous avons proposé dans le Chapitre 8 des couches de V1 et MT évènementielles.

Dans cette implémentation évènementielle, les neurones de V1 sont modélisés comme des neurones intègre-et-tire, où la sortie des filtres de mouvement alimente ces neurone comme un courant externe. La sortie évènementielle des neurones de V1, c'est-à-dire le train de spikes, est transmise directement à la couche suivante de MT. Les neurones de MT sont modélisés comme neurones à conductance intègre-et-tire, où les spikes générés par les neurones de V1 précédents forment la conductance d'entrée excitatrice. D'une façon similaire aux neurones de MT, nous avons aussi implémenté différentes interactions centre-périphérique à caractère inhibiteur. La conductance d'entrée inhibitrice est formée par les spikes générés pour les neurones de V1 localisés dans la région inhibitrice du champ récepteur du neurone de MT.

La principale différence avec l'implémentation analogique des neurones de V1 et MT, est le fait qu'aucune fonction non-linéaire n'a été ajoutée à la sortie de chaque neurone. C'est principalement parce que le processus de production des spikes est déjà non-linéaire que le taux de décharge moyen peut être directement estimé à partir du nombre de spikes émis dans une fenêtre de temps.

11.1.4 La reconnaissance d'actions

L'introduction des modèles sequentiels bio-inspirés pour V1 et MT était motivée par les applications en reconnaissance d'actions pour les scènes réelles.

Pour ce faire, nous avons défini des vecteurs caractéristiques *feature vectors* représentant les séquences d'entrée à partir de la sortie de notre système, c'est à dire, à partir de la sortie des neurones de MT. Ces vecteurs sont utilisés pour la reconnaissance.

En se basant sur l'architecture analogique de V1 et MT, nous avons proposé une carte de mouvement moyen *mean motion map* comme une représentation de

l'information du mouvement contenu dans le stimulus d'entrée. Ce *mean motion map* se définit comme un vecteur d'une longeur égale au nombre total de neurones de MT. Chaque indice du vecteur contient le taux de décharge moyen estimé de chaque neurone de MT à l'intérieur d'une fenêtre temporelle. Avec ces *mean motion maps*, nous avons aussi évalué l'effet de différentes géometries de champs récepteurs de neurone de MT sur cette application de reconnaissance d'actions.

En se basant sur l'architecture évènementielle de V1 et MT, nous avons proposé deux cartes de vitesses *motion maps* différentes, chacune représentant différentes caractérisations des trains de spikes: la carte de mouvement moyen *mean motion map* et la carte de mouvement de synchronie *synchrony motion map*. D'une manière similaire au *mean motion map* défini pour l'architecture analogique de V1 et MT, la *mean motion map* de l'architecture évènementielle est formé à partir du calcul du taux moyen de décharge de chaque neurone de MT dans une fenêtre temporelle. La *synchrony motion map* ne code pas le taux moyen de décharge des neurones de MT, mais leur synchronie. Cette carte se représente avec un tableau de matrices où chaque matrice code la synchronie entre deux neurones de MT de la même couche. Nous avons implementé autant de couches que le nombre des interactions centre-périphérie et *direction-selectivity*. La valeur de la synchronie est calculée en utilisant la représentation ISI et la métrique définie par Kreuz et al. (2007).

11.1.5 Le problème d'ouverture

La deuxième application traitée dans le cadre de cette thèse, est l'étude de l'effet de la suppression périphérique des cellules de V1 dans le problème d'ouverture. En d'autres termes, comment l'information 2D provenant de la suppression périphérique de V1 peut être intégrée par un neurone de MT pour résoudre le problème d'ouverture?

La suppression périphérique, modélisée par une différence de Gaussiennes isotropes (DoG), agit comme une cellule de type *end-stopping* (Sceniak et al. (2001); Jones et al. (2001)). Les cellules *end-stopping* ressortent les réponses des neurones de V1 placés, par exemple, sur les bords d'un stimulus de type barberpole, et inhibent l'activation des cellules de V1 placées au centre.

Pour ce faire, nous avons testé différents stimuli psychophysiques comme les barberpoles, les plaids type I, les plaids type II et les plaids unicinétiques. Grâce à ces stimuli, nous avons observé l'effet d'une suppression périphérique de V1 retardée dans la direction préférée d'un neurone de MT.

11.2 DISCUSSION

Dans cette section, notre objectif est de faire une analyse critique des concepts et des résultats présentés dans cette thèse, mais aussi de discuter des prologements de

ce travail pour les recherches futures.

11.2.1 Modélisation de V1 et MT

Dans V1, la détection du mouvement est faite par des neurones sélectifs à la direction, qui peuvent être trouvés dans les cellules simples et complexes de V1. La propriété de sélectivité à la direction est spécialement renforcée dans les neurones de MT, qui ont également tendance à avoir une vitesse privilégiée par rapport aux cellules de V1. Parmi les cellules de V1 complexes il est possible de trouver des neurones avec des réponses spatio-temporelles indépendantes de la fréquence spatiale du stimulus d'entrée (Priebe et al. (2006)). Ce type de neurones a été également trouvé dans MT (Priebe et al. (2003); Perrone and Thiele (2001)). Ces neurones, également connus comme neurones *speed-tuned*, ont été la cible de différents modèles de V1 et MT proposés dans la littérature (par example, Grzywacz and Yuille (1990); Simoncelli and Heeger (1998); Perrone (2004)).

Dans notre cas, nous n'avons pas modélisé les neurones de type *speed-tuned*. Nos neurones de V1, qui sont basés sur les détecteurs de mouvement proposés par Adelson and Bergen (1985), sont réglés pour une certaine vitesse mais dans une bande spatio-temporelle très limitée. Dans notre cas, les neurones de MT n'utilisent pas les neurones de V1 pour obtenir un plan réglé pour une vitesse donnée, comme l'on fait Simoncelli and Heeger (1998) et Grzywacz and Yuille (1990). Nos neurones de MT groupent les neurones de V1 de la même direction de mouvement préferé, afin d'obtenir une réponse sensible à une direction donnée pour toutes les fréquences spatiotemporelles de l'espace. Ce mécanisme de groupement simplifié a été inspiré pour l'application de la reconnaissance d'actions, où des études comme celles de Casile and Giese (2005), ont montré que notre capacité de reconnaissance nécessite peu d'information sur la vitesse et sur la localisation spatialle du mouvement en comparaison à la direction du mouvement.

▼

Une extension naturelle de notre travail est de considérer l'implementation des neurones de type ***speed-tuned****. Ce type de neurones est nécessaire pour l'implémentation d'un système qui manipule d'une manière plus appropriée certains phénomènes observés dans V1 et MT, par exemple, les plaids type II et les neurones de type* ***pattern-component****.*

Détection de mouvement dans V1

Pour la détection rapide de mouvement local, Simoncelli and Heeger (1998) ont proposé des entités locales modélisées comme des filtres d'énergie spatio-temporels. Toutefois, ces filtres ont un profil temporel qui n'est pas causal ce qui est contradictroire avec la physiologie des neurones de V1. Notre approche, d'autre part, implémente des profils temporels en accord avec la physiologie des neurones de V1.

Ces profils spatio-temporels biologiquement inspirés nécessitents des calculs difficiles pour le réglage de l'orientation spatiotemporelle dans l'espace des fréquences. En conséquence, le réglage de l'orientation du mouvement doit être calculé par approximation numérique.

Mécanisme de groupement de MT

Dans cette thèse, nous avons d'abord présenté un neurone de MT comme une entité qui groupe les sorties des cellules complexes de V1. Nous montrons que cette procedure de groupage donne comme résultat des neurones de MT avec une haute sélectivité à la direction, propriété très importante dans les neurones de MT réels (Albright (1984); Snowden et al. (1991); Lagae et al. (1993); Churchland et al. (2005)), mais comme nous avons déjà discuté, nous n'obtenons pas des neurones réglés pour une certaine vitesse. L'activité dynamique des neurones de MT dépend du modèle implémenté pour les neurones de V1. Par exemple, dans le Chapitre 9 les neurones de MT groupent seulement les taux de décharge moyens des neurones de V1. Dans le Chapitre 8, par contre, les neurones de MT intègrent les *spikes* générés par les neurones de V1, tant que dans le Chapitre 7, les neurones de MT intègrent les taux de décharge moyens des neurones de V1 qui sont à valeurs analogiques.

Interactions centre-périphérique

La plus grande partie des cellules simples et complexes de V1 ont des interactions centre-périphérie, qui peuvent être intégratives ou suppressives selon les propriétés du stimulus d'entrée, par exemple, le contraste, la fréquence spatio-temporelle, le type de stimulus (des bars, des points aléatoires, des plaids, etc). D'une manière similaire, les cellules de MT ont montré aussi des interactions centre-périphérique dépendant aussi du stimulus d'entrée. On pense que le mécanisme de la suppression périphérique est impliqué dans la solution du problème de l'ouverture modulant la direction préferée des neurones de MT.

Nous avons présenté différentes interactions centre-périphérique pour les neurones de MT qui sont implémentées dans les Chapitres 7 et 8. Nous avons montré que cette interaction, qui est le plus souvent suppressive, code des singularités importantes de l'information du mouvement qui seront cruciales pour des applications réelles, comme la reconnaissance d'actions. Nous avons implémenté aussi des interactions centre-périphérique pour les neurones de V1 (Chapitre 9) en montrant que la suppression périphérique des cellules de V1 peut faire sortir l'information de mouvement 2D et contribuer à la solution du problème de l'ouverture. Nous n'avons pas essayé différentes configurations de périphérie pour les neurones de V1.

Notre modèle implémente différents champs récepteurs de cellules de MT non-classiques, avec différentes interactions centre-périphérique (par exemple, Xiao et al. (1995), Born (2000)). Le rôle de ces différentes formes de champs récepteurs de neu-

rones de MT dans la reconnaissance d'actions n'avait pas été encore évalué. Dans le cadre de cette thèse nous montrons des résultats pour la performance de la reconnaissance d'actions en utilisant différentes structures et géometries des champ récepteurs comme ceux observés dans l'aire MT chez le macaque (Born (2000); Xiao et al. (1995, 1997b)), et en montrant leur rôle important pour la répresentation du mouvement.

Feedbacks?

La citation au début du Chapitre 5[1] a vraiment inspiré notre travail et elle explique aussi pourquoi nous n'avons pas implémenté des feedbacks dans notre modèle.

▼

L'analyse du mouvement peut être réalisée sans l'implémentation de feedbacks, mais bien sûr, les feedbacks peuvent ajouter à notre modèle des capacités additionelles et de la robustesse. L'implémentation des feedbacks[2] *serait donc une perspective naturelle. Par exemple, notre modèle n'est pas capable de traiter la foule, ou il aura des problèmes pour traiter les occlusions ou arrière-plans complexes, tâches qui ont été aussi implémentées au travers des mécanismes de l'attention. Pour le cas précis de la résolution du problème d'ouverture, les feedbacks jouent aussi un rôle important, par exemple, en diffusant l'information non-ambiguë obtenue par les couches supérieures sur les neurones des couches inférieures pour faire sortir la direction réelle du mouvement des objets et ainsi résoudre le problème d'ouverture.*

Vers la modèlisation du système visuel?

Dans le cadre de cette thèse, nous nous sommes concentrés sur une petite partie du système visuel (voir le schéma de Felleman & Van Essen montré dans l'introduction). Nous imaginons ce travail comme une contribution à une modèlisation plus large du système visuel.

▼

Par exemple, notre modèle s'occupe de l'analyse du mouvement en considérant que les images d'entrée arrivent directement sur V1, où une première étape de filtrage est exécutée. Mais cette simplification manque une partie du système visuel importante à savoir, tout le chemin entre la rétine et V1. Donc, la première idée qui vient à l'esprit est de connecter notre système à la sortie d'un modèle de rétine (comme ceux proposés, par example, par Hérault and Durette (2007); Wohrer and Kornprobst (2009)). Quel serait le gain de cette nouvelle architecture? Peut-être un modèle simplifié de rétine

[1] *"We cannot think about what a feedback interaction could do if we do not first explore the limitations of a feedforward model"* – Simon Thorpe (GDR-vision meeting 2008)

[2] Nous faisons la remarque que l'implémentation des connections de type feedbacks sont déjà implémentés dans plusieurs modèles de traitement du mouvement (voir le Chapitre 4). Cependant, leur implémentation dans notre modèle pourrait être intéressante pour étudier leur rôles dans le traitement et l'intégration du mouvement.

exécutant certaines opérations comme la détection de bords et la normalisation du contraste, pourrait être suffisants?

D'une façon plus générale, une perspective ambitieuse est de considérer notre contribution comme une partie d'un modèle du système visuel global. L'intégration de différents élements du système visuel requièrent de gros efforts pour comprendre et implémenter la connectivité entre les différentes aires et couches. Un gros effort dans cette direction est mené dans le BlueBrainProject[3]*, qui s'interesse à la modélisation des colonnes souscorticales pour les utiliser dans la simulation des aires du cerveau.*

Pour affronter cette perspective à long terme, plusieurs questions difficiles devront être considerées: Quel type de cadre mathématique serait le plus appropié? Comment pourrions nous traiter la fusion des différentes échelles? Quels sont les challenges computationnels pour l'implémentation?

11.2.2 Reconnaissance d'actions: l'analyse des résultats

Une architecture analogique ou évènementielle?

Avec l'architecture analogique, nous avons montré que l'inclusion de différentes interactions centre-périphérie des neurones de MT peut améliorer significativement les performance de la reconnaissance. Nous pensons que les différentes interactions centre-périphérie des neurones de MT extraient des échantillons singuliers de mouvement qui se comportent comme de l'information clé pour une tâche de catégorisation du mouvement.

Avec l'achitecture évènementielle, les résultats de la reconnaissance obtenus à travers des *synchrony motion maps* sont légèrement inférieurs à ceux obtenus avec des *mean motion maps*, particulièrement si nous considérons seulement l'activation du champ récepteur classique des neurones de MT. Cette différence est mise en valeur dans les expériences de robustesse. Comme une possible explication, nous croyons que l'analyse de la synchronie oublie largement le taux moyen de décharge, qui répresente information fondamentale de l'activation du réseau. Néanmoins, avec seulement l'information obtenue par l'analyse de la synchronie, la performance de la reconnaissance peut être satisfaisante. Également, remarquons que l'analyse de la synchronie pour coder l'information du mouvement améliore la séparabilité entre-classes obtenant de cette manière un meilleur partitionement *clustering* (voir la Figure 8.9 et Tableau 8.3). Ces résultats sont en accord avec des mesures réelle par rapport à l'information complémentaire du codage par le taux moyen de décharge ou par la synchronie: Il y a des preuves que le système moteur et le cortex cérébral visuel utilisent tous les deux le taux moyen de décharge et la synchronie comme une analyse conjointe pour extraire l'information complémentaire (Maldonado et al. (2008); Grammont and Riehle (2003); Riehle et al. (1997)). Comme perspective nous pensons

[3] `http://bluebrain.epfl.ch`

combiner les deux cartes de mouvement proposées dans cette thèse pour réusir une meilleure répresentation de l'information du mouvement d'entrée.

Bien que les *mean motion maps* des architectures analogiques et évènementielles aient la même philosophie de construction, il y a des différences dans les taux de reconnaissance obtenus pour chaque architecture. Nous croyons que cette différence est due au fait que les paramètres du modèle ne sont pas les mêmes et que les mécanismes de production des spikes pour les deux architectures ne sont pas équivalents.

D'une une manière générale, les résultats obtenus avec l'architecture évènementielle ne sont pas aussi satisfaisants que ceux obtenus avec l'architecture analogique. Cela ne signifie pas que l'utilisation des spikes n'ouvre pas de nouvelles perspectives, mais cela veut plutôt dire que notre interpretation actuelle du codage neuronal n'est sans doute pas appropriée ou que nos cartes de vitesses *motion maps* proposées comme réprésentations de l'information du mouvement d'entrée doivent être considerées comme une répresentation imparfaite des données.

▼

À propos de l'architecture évènementielle, notre principale perspective est l'étude des ***statistiques d'ordre supérieur pour l'analyse des trains des spikes*** *générés par les neurones de MT. Il est probable que cette étude puisse nous donner de nouvelles idées par rapport à l'analyse de la sortie de MT, c'est-à-dire dans notre application, le contenu en mouvement dans les vidéos d'entrée.*

Plus des validations?

Bien sûr, plus de validations sont aussi nécessaires. Nous avons testé notre modèle avec la base de données de Weizmann. La bonne performance obtenue avec notre modèle, tant que pour l'architecture analogique, que pour l'architecture évènementielle, montre le bien fondé de la répresentation de nos cartes de mouvement.

La base de données de Weizmann a été aussi utilisée par Blank et al. (2005) et Jhuang et al. (2007) pour valider leurs approches. Cependant, les conditions de tests et le protocol d'expérimentation ne sont pas les mêmes que ceux considérés dans nos expériences, donc les résultats pour la reconnaissance ne peuvent pas être comparés d'une manière directe.

À partir de nos résultats, nous ne prétendons pas dire que notre modèle fonctionne dans toutes les conditions. Mais cette réserve est en realité générale comme Pinto et al. (2008) l'ont bien remarqué: Déclarer que le problème global de la reconnaissance d'actions est résolu à partir des résultat obtenus pour une seule base de données n'est pas réaliste. Conditions réelles comme les arrières-plans complexes, lesrotations, les occlusions ou les distracteurs sont normalement pas inclus dans les bases de données. Pour aller plus loin, des validations avec bases de données différentes (comme celle du KTH[4]) sont absolument nécessaires.

[4] `http://www.nada.kth.se/cvap/actions/`

▼

En plus de mener davantage de validations, nous pensons à deux autres perspectives. La première est l'étude de comment l'information provenant de la ***forme*** *peut être ajoutée à notre modèle d'une manière dynamique et ainsi l'intégrer à nos cartes de mouvement, particulièrement dans le cas des arrières-plans complexes où cette intégration pourrait jouer un rôle important. La deuxième perspective est l'étude du* ***rôle des différents classificateurs*** *dans nos résultats. La qualité de la représentation de nos cartes de mouvement affecte clairement la performance dans la reconnaissance. Mais, le choix du classificateur est aussi un élement crucial dans notre système. Comme perspective, il serait aussi intéressant de faire un test de performance (benchmark) entre différents classificateurs pour ainsi évaluer leur impact sur la performance de la reconnaissance d'actions.*

Nous avons aussi pensé à d'autres expériences à réaliser. Par exemple, nous nous sommes interrogés sur l'existence de ***vues clés*** *(keyframes). Comme il a été déjà demontré, toutes les vues d'une séquence d'action n'ont pas le même intérêt pour la reconnaissance. Si seulement quelques vues clés sont montrées, nous pouvons tout de même réaliser la reconnaissance d'actions sans difficulté. Il serait intéressant d'implémenter, selon la complexité de l'information du mouvement de chaque séquence, une extraction automatique des vues clé pour ainsi comparer seulement l'information la plus utile du mouvement dans chaque séquence. Une autre série d'expériences serait liée à la* ***reconnaissance rapide****. Suite aux expériences de reconnaissance rapide (ultra fast categorization), effectuées notamment par Thorpe et al. (2001) et Thorpe and Fabre-Thorpe (2001), nous pourrions étudier l'évolution de la performance de la reconnaissance au cours du temps. En d'autres termes, combien de temps est requis pour avoir une performance satisfaisante (taille de la fenêtre de temps, périodes d'echantillonage* Δt*)? Ces études pourraient être considérées comme une extension du* ***rank-order-coding*** *à l'analyse vidéo. Les changements de la valeur de* Δt *pourraient être aussi interprétés en lien avec la "mémoire" du système, c'est-à-dire, combien d'information dans le passé doit être prise pour avoir une bonne représentation de chaque action?*

Comment comparer notre modèle avec l'existant?

Des modèles antérieurs ont permis de suggérer que la perception du mouvement biologique dépend de fortes interactions entre le chemin dédié au traitement du mouvement et le chemin dédié au traitement de la forme (Blake and Shiffrar (2007)). Dans le modèle proposé par Giese and Poggio (2003), le chemin du mouvement comme le chemin de la forme apprennent des séquences ou certaines vues clé (snapshots) de formes humaines et de pattern de flux-optique, respectivement. D'autre part, l'information spatiale de ces patterns de mouvement est importante pour la reconnaissance du mouvement biologique a été étudié par exemple par Hiris et al. (2005). Plusieurs modèles ont été proposés pour contraindre dynamiquement l'information du mouvement selon l'information locale de la forme (voir par exam-

ple, Bayerl and Neumann (2007); Tlapale et al. (2008)).

Plus spécialement, Giese and Poggio (2003) ont proposé un modèle neurophysiologique pour le traitement de l'information visuelle dans la voie dorsale (*mouvement*) et ventrale (*forme*). Le modèle a été validé pour la reconnaissance d'actions utilisant comme stimulus d'entrée des figures *batôn* construites à partir des séquences réelles. En supposant qu'il n'y a pas d'interaction entre les deux voies, les auteurs ont trouvé que chaque voie indépendamment est capable d'effectuer la reconnaissance d'actions. De plus, leur modèle présente des propriétés très intéressantes pour la reconnaissance de mouvements biologiques, comme l'invariance à l'échelle spatiale et temporelle, la robustesse au bruit ajouté aux stimuli d'entrée, etc. L'une des différences plus importantes avec notre approche est que plusieurs paramètres doivent être ajustés pour considérer une nouvelle action. Dans notre cas, nous n'avons pas besoin de ajuster des paramètres pour tenir compte d'une nouvelle action. Il suffit d'insérer les nouvelles cartes de mouvement dans le *training set*.

Dans le travail plus récent de Jhuang et al. (2007), les auteurs implementent aussi une invariance à l'échelle spatiale et temporelle (c'est-à-dire, à la taille du stimulus et la durée de l'action, respectivement). Leur approche utilise un modèle bio-inspiré pour la reconnaissance d'actions, inspiré par Giese and Poggio (2003) et Serre et al. (2005). L'invariance à l'échelle spatiale et temporelle est obtenue en choisissant autant de couches de détecteurs de mouvement que d'échelles spatio-temporelles à détecter, suivi par un operateur *max*. Ces méchanismes peuvent facilement être implémentés en ajoutant davantage de couches avec différentes échelles spatio-temporelles, suivies par l'operateur *max* entre les différentes couches qui codent la même direction de mouvement.

Contrairement aux modèles de flot-optique, où une seule valeur de vitesse est associée à chaque point, notre modèle reproduit la richesse des interactions centre-périphérie des cellules de MT en variant les géometries des périphéries (voir Figure 5.14 et Born (2000); Xiao et al. (1995, 1997b)). Les différentes géométries des périphéries donnent différents types de contrastes de mouvement, pour plusieurs orientations en chaque point. De façon intéressante, nous avons montré qu'en ajoutant cette diversité des cellules de MT, la performance de la reconnaissance d'action est améliorée. Notre interprétation est que les cellules avec périphéries inhibitrices amènent une information reliée aux contrastes de vitesses ou singularités dans le champ de vitesse du stimulus d'entrée.

Concernant une comparaison quantitative, nous avons suivi le protocole experimental presenté par Jhuang et al. (2007) pour comparer leurs résultats avec les notres. Pour nos deux architectures, les résultats obtenus avec *mean motion maps* et *synchrony motion maps* ont montré une haute variabilité selon les séquences considérées dans le *training set*. À cause de cette haute variabilité trouvé edans nos résultats, la comparaison directe avec Jhuang et al. (2007) n'est pas évidente, et les pourcentages de reconnaissance ne sont pas représentatifs.

11.2.3 La suppression périphérique des cellules de V1: analyse de résultats

Comme nous l'avons déjà mentionné dans le Chapitre 9, nous avons obtenu un petit décalage dans la direction préférée des neurones de MT par rapport aux valeurs rapportées par Pack et al. (2004). Par conséquent, nous avons besoin d'ajouter des mécanismes supplémentaires pour vraiment résoudre le problème de l'ouverture. Pour obtenir une solution satisfaisante au problème d'ouverture, plusieurs solutions pourraient être analysées:

- Les changements de la direction préférée d'un neurone de MT dépend de plusieurs paramètres au cours d'une étape non-linéaire, à la sortie du traitement de V1. Cette dépendance est cohérente avec les résultats rapportés par Rust et al. (2006) exprimant que la sélectivité des neurones de MT de type "pattern" dépend fortement du traitement non-linéaire (par exemple, avec la normalisation divisive) autant au niveau de V1 que de MT. Les non-linéarités jouent apparemment un rôle fondamental qui doit être clarifié.

- Différent types d'interactions centre-périphérie en travaillant en différent échelles spatiales et avec différentes orientations relatives pourraient être une alternative à l'extraction de caractéristiques 2D du mouvement.

- Un meilleur processus de diffusion de l'information 2D pourrait être réalisé en utilisant des interactions anisotropes entre neurones placés à différents endroits du champ visuel (voir par exemple Tlapale et al. (2008)).

- L'information 2D extraite par le mécanisme de suppression périphérique dans les neurones de V1 pourrait être aussi étudiée dans ce cadre. Dans le cas des plaids type II ou plaids unicinétiques, le mouvement perçu pourrait être associé à la détection du mouvement de caractéristiques 2D qui ont une fréquence spatiale plus basse que les gratings utilisés pour le stimulus (Wilson et al. (1992)). Le réglage spatiotemporel des neurones de V1 utilisés pour l'extraction de l'information 2D est crucial pour détecter ce type de caractéristiques, et pour avoir aussi une forte réponse aux caractéristiques 2D.

Nous avons aussi exploré l'effet de la suppression périphérique des neurones de V1 pour différents stimuli comme les plaids de type I, les plaids de type II et les plaids unicinétiques. Le seul plaid pour laquel un déplacement dans la direction préférée d'un neurone de MT a été perçu, est le cas du plaid unicinétique, avec un déplacement d'environ 10°. Nous attendions un déplacement similaire pour les plaids de type II, mais dans nos expériences la suppression périphérique des neurones de V1 n'a pas affecté la direction préférée des neurones de MT. Nous croyons que cet effet est relié à la précision de la fréquence spatio-temporelle qui doit être "vu' par nos détecteurs de mouvement de V1, où le réglage fréquentiel des neurones de V1 est essentiel pour extraire de manière correcte les caractéristiques du mouvement 2D.

▼

Le modèle des neurones de V1 défini dans le Chapitre 9 peut être formalisé comme un modèle de masse neurale où plusieurs résultats théoriques et expérimentaux sont déjà établis (voir par exemple, Faugeras et al. (2009); Giese (1998)). En fait, l'équation (9.1) *est équivalente au modèle à voltage de masse neurale, où la population des neurones dans notre cas serait la combination de quatre neurones simples (voir équation* (5.9)*). L'implémentation du modèle de masse neurale requiert une analyse mathématique plus profonde: Il faut étudier les solutions stationnaires, la stabilité, le diagramme de bifurcations, etc. Par exemple, il serait intéressant d'étudier si le nombre des solutions change (multistabilité) avec différentes configurations de centre-périphérie des neurones de V1 ou MT.*

11.2.4 Contribution logiciele

Un effort considérable a été mené pour implémenter les méthodes nécessaires afin de construire les réseaux de neurones, les neurones évènementiels, le filtrage du mouvement basé sur l'energie, etc...

En particulier, nous avons chercher à bien comprendre et optimiser le filtrage spatio-temporel nécessaire au calcul du mouvement basé sur l'énergie. Cette étape représente le calcul le plus exigeant. Cette partie du travail sera bientôt disponible comme une bibliothèque libre en C/C++.

L'implémentation des neurones évènementiels a été faite grâce à la bibliothèque MVAspike développée par Rochel (2004). Cette bibliothèque nous a permi de créer des couches de neurones évènementiels et de les relier.

CHAPTER 12

PUBLICATIONS ARISING FROM THIS WORK

Journal papers

1. Maria-Jose Escobar, Guillaume S. Masson, Thierry Vieville, Pierre Kornprobst. *Action Recognition Using a Bio-Inspired Feedforward Spiking Network*. International Journal of Computer Vision (IJCV), Volume 82, Number 3, Pages 284–301, 2009.

Conference papers

1. Maria-Jose Escobar, Pierre Kornprobst. *Action Recognition with a Bio-Inspired Feedforward Motion Processing Model: The Richness of Center-Surround Interactions*. Computer Vision - ECCV, Lecture Notes in Computer Science, pages 186-199, 2008.

2. Maria-Jose Escobar, Guillaume S. Masson and Pierre Kornprobst. *A Simple Mechanism to Reproduce the Neural Solution of the Aperture Problem in Monkey Area MT*. Neurocomp 2008.

3. Maria-Jose Escobar, Thierry Vieville and Pierre Kornprobst. *Biological Motion Recognition Using a MT-like Model*. Neurocomp 2006.

Conference abstracts

1. Maria-Jose Escobar, Guillaume S. Masson, Thierry Vieville and Pierre Kornprobst. *Spiking MT model: Dynamics and motion patterns*. European Conference in Visual Perception (ECVP), 2007.

2. Maria-Jose Escobar, Thierry Vieville and Pierre Kornprobst. *Spike to Spike Model and Applications*. Computational Neuroscience Meeting (CNS), 2007.

3. Maria-Jose Escobar, Adrien Wohrer, Pierre Kornprobst and Thierry Vieville. *Can we recognize motion from spike train analysis?*. European Conference in Visual Perception (ECVP), 2006.

Research Reports

1. Maria-Jose Escobar, Guillaume S. Masson, Thierry Vieville and Pierre Kornprobst. *Rate Versus Synchrony Code for Human Action Recognition*. INRIA Research Report RR-6669, 2008.

2. Maria-Jose Escobar, Guillaume S. Masson and Pierre Kornprobst. *A Simple Mechanism to Reproduce the Neural Solution of the Aperture Problem in Monkey Area MT*. INRIA Research Report RR-6579, 2008.

3. Maria-Jose Escobar, Guillaume S. Masson, Thierry Vieville and Pierre Kornprobst. *Spike to Spike Model and Applications: A biological plausible approach for the motion processing*. INRIA Research Report RR-6280, 2007.

Bibliography

Adelson, E. and J. Bergen: 1985, 'Spatiotemporal energy models for the perception of motion'. *Journal of the Optical Society of America A* **2**, 284–299. [47, 48, 50, 51, 53, 58, 82, 84, 161, 164, 173, 177]

Adelson, E. and J. Bergen: 1986, 'The extraction of Spatio-temporal Energy in Human and Machine Vision'. In: *Workshop on Motion : Representation and Analysis*. pp. 151–155. [51]

Adelson, E. and J. Movshon: 1982, 'Phenomenal coherence of moving visual patterns'. *Nature* **300**(5892), 523–525. [57]

Aggarwal, J. and Q. Cai: 1999, 'Human motion analysis: a review'. *Computer Vision and Image Understanding* **73**(3), 428–440. [101]

Albrecht, D. G., W. S. Geisler, and A. M. Crane: 2004, 'Nonlinear properties of visual cortex neurons: Temporal dynamics, stimulus selectivity, neural performance.'. In: L. M. Challupa and J. S. Werner (eds.): *The Visual Neurosciences*, Vol. 1. MIT press, pp. 747—764. [111]

Albright, T. and R. Desimone: 1987, 'Local precision of visuotopic organization in the middle temporal area (MT) of the macaque'. *Experimental Brain Research* **65**(3), 582–592. [32]

Albright, T. D.: 1984, 'Direction and orientation selectivity of neurons in visual area MT of the macaque'. *Journal of Neurophysiology* **52**(6), 1106–1030. [32, 165, 178]

Alonso, J., W. Usrey, and R. Reid: 2001, 'Rules of connectivity between geniculate cells and simple cells in cat primary visual cortex'. *The Journal of Neuroscience* **21**(11), 4002–4015. [23]

Angelucci, A. and J. Bullier: 2002, 'Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons?'. *J. Physiol. (Paris)*. [30]

Angelucci, A. and J. Bullier: 2003, 'Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons?'. *J Physiol Paris* **97**(2–3), 141–154. [30]

Angelucci, A., J. Levitt, E. Walton, J. Hupe, J. Bullier, and J. Lund: 2002, 'Circuits for local and global signal integration in primary visual cortex'. *The Journal of Neuroscience* **22**(19), 8633–8646. [29]

Bair, W., J. R. Cavanaugh, and A. Movshon: 2003, 'Time Course and Time–Distance Relationships for Surround Suppression in Macaque V1 Neurons'. *The Journal of Neuroscience* **23**(20), 7690–7701. [29, 30]

Barron, J., D. Fleet, and S. Beauchemin: 1994, 'Performance of Optical Flow Techniques'. *The International Journal of Computer Vision* **12**(1), 43–77. [45]

Basole, A., L. White, and D. Fitzpatrick: 2003, 'Mapping multiple features in the population response of visual cortex'. *Nature* **423**(6943), 986–990. [80]

Bayerl, P.: 2005, 'A model of visual motion perception'. Ph.D. thesis, Ulm University. [74]

Bayerl, P. and H. Neumann: 2004, 'Disambiguating Visual Motion Through Contextual Feedback Modulation'. *Neural Computation* **16**(10), 2041–2066. [vi, 44, 52, 74, 76]

Bayerl, P. and H. Neumann: 2005, 'Attention and figure-ground segregation in a model of motion perception'. *Journal of Vision* **5**(8), 659–659. [74]

Bayerl, P. and H. Neumann: 2007, 'Disambiguating Visual Motion by Form–Motion Interaction – a Computational Model'. *International Journal of Computer Vision* **72**(1), 27–45. [38, 74, 151, 169, 183]

Beintema, J. and M. Lappe: 2002, 'Perception of biological motion without local image motion'. *Proceedings of the National Academy of Sciences of the USA* **99**(8), 5661–5663. [103]

Bergen, J. R. and M. S. Landy: 1991, 'Computational Modeling of Visual Texture Segregation'. In: M. Landy and A. Movshon (eds.): *Computational Models of Visual Processing*. MIT press, pp. 253–271. [69]

Berzhanskaya, J., S. Grossberg, and E. Mingolla: 2007, 'Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception'. *Spatial Vision* **20**(4), 337–395. [38, 73, 150, 151]

Biederlack, J., M. Castelo-Branco, S. Neuenschwander, D. W. Wheeler, W. Singer, and D. Nikolić: 2006, 'Brightness induction: rate enhancement and neuronal synchronization as complementary codes.'. *Neuron* **52**(6), 1073–1083. 07042. [128]

Black, M.: 1992, 'Robust incremental optical flow'. Ph.D. thesis, Yale University, Department of Computer Science. [47]

Black, M. and P. Rangarajan: 1996, 'On the unification of line processes, outlier rejection, and robust statistics with applications in early vision'. *The International Journal of Computer Vision* **19**(1), 57–91. [47]

Blake, R. and M. Shiffrar: 2007, 'Perception of Human Motion'. *Annual Review of Psychology* (58), 12.1–12.27. [103, 169, 182]

Blank, M., L. Gorelick, E. Shechtman, M. Irani, and R. Basri: 2005, 'Actions as Space-Time Shapes'. In: *Proceedings of the 10th International Conference on Computer Vision*, Vol. 2. pp. 1395–1402. [101, 102, 124, 142, 168, 181]

Bobick, A. and J. Davis: 2001, 'The recognition of human movement using temporal templates'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(3), 257–267. [101]

Born, R. and D. Bradley: 2005, 'Structure and Function of Visual Area MT'. *Annu. Rev. Neurosci* **28**, 157–189. [31]

Born, R. T.: 2000, 'Center-Surround Interactions in the Middle Temporal Visual Area of the Owl Monkey'. *Journal of Neurophysioly* **84**, 2658–2669. [37, 39, 92, 134, 166, 170, 178, 179, 183]

Born, R. T., C. C. Pack, C. Ponce, and S. Yi: 2006, 'Temporal Evolution of 2–Dimensional Direction Signals Used to Guide Eye Movements'. *Journal of Neurophysiology* **95**, 284–300. [38, 148]

Borst, A.: 2007, 'Correlation versus gradient type motion detectors: the pros and cons'. *Philosophical Transactions of the Royal Society of London: Series B, biological sciences* **362**(1479), 369–374. [45, 53]

Buracas, G. T. and T. D. Albright: 1996, 'Contribution of area MT to perception of three-dimensional shape: a computational study.'. *Vision Res* **36**(6), 869–87. [36]

Carandini, M., J. B. Demb, V. Mante, D. J. Tollhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust: 2005, 'Do we know what the early visual system does?'. *Journal of Neuroscience* **25**(46), 10577–10597. [24]

Casile, A. and M. Giese: 2003, 'Roles of motion and form in biological motion recognition'. *Artifical Networks and Neural Information Processing, Lecture Notes in Computer Science 2714* pp. 854–862. [104]

Casile, A. and M. Giese: 2005, 'Critical features for the recognition of biological motion'. *Journal of Vision* **5**, 348–360. [104, 164, 177]

Cessac, B., H. Rostro, J.-C. Vasquez, and T. Viéville: 2008, 'Statistics of spikes trains, synaptic plasticity and Gibbs distributions,'. In: *proceedings of the conference NeuroComp 2008 (Marseille)*. [128]

Chapman, B., K. Zahs, and M. Stryker: 1991, 'Relation of cortical cell orientation selectivity to alignment of receptive fields of the geniculocortical afferents that arborize within a single orientation column in ferret visual cortex'. *Journal of Neuroscience* **11**(5), 1347–1458. [23]

Chey, J., S. Grossberg, and E. Mingolla: 1997, 'Neural dynamics of motion processing and speed discrimination'. *Vision Res.* **38**, 2769–2786. [vi, 71, 72]

Chomat, O., J. Martin, and J. L. Crowley: 2000, 'A probabilistic sensor for the perception and recognition of activities'. In: *Proceedings of the 6th European Conference on Computer Vision*, Vol. 1842. Dublin, Ireland: Springer Berlin / Heidelberg, pp. 487–503. [102]

Churchland, A. and S. Lisberger: 2005, 'Discharge properties of MST neurons that project to the frontal pursuit area in macaque monkeys'. *The Journal of Neurophysiology* **94**(2), 1084–1090. [41]

Churchland, M. M., N. J. Priebe, and S. G. Lisberger: 2005, 'Comparison of the Spatial Limits on Direction Selectivity in Visual Areas MT and V1'. *Journal of Neurophysiology* **93**, 1235—1245. [31, 32, 34, 165, 178]

Collins, R., R. Gross, and J. Shi: 2002, 'Silhouette-based human identification from body shape and gait'. In: *5th Intl. Conf. on Automatic Face and Gesture Recognition*. p. 366. [102]

Conway, B. and M. Livingstone: 2003, 'Space-Time Maps and Two-Bar Interactions of Different Classes of Direction-Selective Cells in Macaque V1'. *Journal of Neurophysiology* **89**, 2726–2742. [25, 48, 81]

Cutler, R. and L. Davis: 2000, 'Robust real-time periodic motion detection, analysis, and applications'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8). [102]

Dayan, P. and L. F. Abbott: 2001, *Theoretical Neuroscience : Computational and Mathematical Modeling of Neural Systems*. MIT Press. [130]

De Valois, R., N. Cottaris, et al.: 2000, 'Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity'. *Vision Research* **40**, 3685–3702. [23, 24, 25, 27, 81, 83]

Deangelis, G. C. and A. Akiyuki: 2004, 'A Modern View of the Classical Receptive Field: Linear and Nonlinear Spatiotemporal Processing by V1 Neurons.'. In: L. M. Challupa and J. S. Werner (eds.): *The Visual Neurosciences*, Vol. 1. MIT press, pp. 704–719. [25, 26]

Derrington, A. M. and B. S. Webb: 2004, 'Visual System: How is the Retina Wired Up to the Cortex?'. *Current Biology* **14**, R14–R15. [24]

Destexhe, A., M. Rudolph, and D. Paré: 2003, 'The high-conductance state of neocortical neurons in vivo'. *Nature Reviews Neuroscience* **4**, 739–751. [112, 131]

Dollar, P., V. Rabaud, G. Cottrell, and S. Belongie: 2005, 'Behavior recognition via sparse spatio-temporal features'. In: *VS-PETS*. pp. 65–72. [101, 103]

Dow, B. M., A. Z. Snyder, R. G. Vautin, and R. Bauer: 1981, 'Magnification factor and receptive field size in foveal striate cortex of the monkey'. *Experimental Brain Research* **44**, 213–228. [32]

Duffy, C. and R. Wurtz: 1991, 'Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli'. *The Journal of Neurophysiology* **65**(5), 1329–1345. [41]

Duffy, C. and R. Wurtz: 1997, 'MST neurons code for visual motion in space independent of pursuit eye movements'. *Journal of Neurophysiology* **97**(5), 3473–3483. [42]

Efros, A., A. Berg, G. Mori, and J. Malik: 2003, 'Recognizing Action at A Distance'. In: *Proceedings of the 9th International Conference on Computer Vision*, Vol. 2. pp. 726–734. [101, 102]

Emerson, R., M. Citron, W. Vaughn, and S. Klein: 1987, 'Nonlinear directionally selective subunits in complex cells of cat striate cortex'. *Journal of Neurophysiology* **58**(1), 33–65. [25]

Enkelmann, W.: 1988, 'Investigation of multigrid algorithms for the estimation of optical flow fields in image sequences'. *Computer Vision, Graphics, and Image Processing* **43**, 150–177. [47]

Erol, A., G. Bebisa, M. Nicolescua, R. D. Boyleb, and X. Twomblyb: 2007, 'Vision-based hand pose estimation: A review'. *Computer Vision and Image Understanding* **108**(1–2), 52–73. [101]

Escobar, M.-J. and P. Kornprobst: 2008, 'Action Recognition with a Bio–Inspired Feedforward Motion Processing Model: The Richness of Center-Surround Interactions'. In: *Proceedings of the 10th European Conference on Computer Vision*, Vol. 5305 of *LNCS*. pp. 186–199, Springer–Verlag. [102]

Fanti, C., L. Zelnik-Manor, and P. Perona: 2005, 'Hybrid models for human motion recognition'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Vol. 1. pp. 1166–1173. [102]

Fathi, A. and G. Mori: 2008, 'Action recognition by learning mid-level motion features'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Faugeras, O., R. Veltz, and F. Grimbert: 2009, 'Persistent neural states: stationary localized activity patterns in nonlinear continuous n-population, q-dimensional neural networks'. *Neural Computation* **21**(1), 147–187. [171, 185]

Felleman, D. and D. Van Essen: 1991, 'Distributed hierarchical processing in the primate cerebral cortex'. *Cereb Cortex* **1**, 1–47. [2, 10, 31, 33]

Fellous, J.-M., P. H. E. Tiesinga, P. J. Thomas, and T. J. Sejnowski: 2004, 'Discovering Spike Patterns in Neural Responses'. *The Journal of Neuroscience* **24**(12), 2989–3001. [128, 130]

Ferrera, V. and H. Wilson: 1990, 'Perceived direction of moving two-dimensional patterns'. *Vision Research* **30**(2), 273–287. [57, 58]

Fleet, D. and A. Jepson: 1990, 'Computation of component image velocity from local phase information'. *The International Journal of Computer Vision* **5**, 77–104. [47]

Fleet, D. J. and A. D. Jepson: 1989, 'Hierarchical Construction of Orientation and Velocity Selective Filters'. *pami* **11**(3), 315–325. [48]

Fleet, D. J. and Y. Weiss: 2005, 'Optical Flow Estimation'. In: N. Paragios, Y. Chen, and O. Faugeras (eds.): *Mathematical Models for Computer Vision: The Handboook*. Springer. [45]

Fries, P., S. Neuenschwander, A. K. Engel, R. Goebel, and W. Singer: 2001, 'Rapid feature selective neuronal synchronization through correlated latency shifting.'. *Nat Neurosci* **4**(2), 194–200. 07045. [128]

Gautrais, J. and S. Thorpe: 1998, 'Rate Coding vs Temporal Order Coding : a theorical approach'. *Biosystems* **48**, 57–65. [127]

Gavrila, D.: 1999, 'The visual analysis of human movement: A survey'. *Computer Vision and Image Understanding* **73**(1), 82–98. [101, 102]

Gavrila, D. and L. Davis: 1996, '3-D Model-based Tracking of Humans in Action: a Multi-view Approach'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, IEEE. [102]

Geesaman, B. and R. Andersen: 1996, 'The analysis of complex motion patterns by form/cue invariant MSTd neurons'. *The Journal of Neuroscience* **16**(15), 4616–4632. [42]

Gerstner, W. and W. Kistler: 2002, *Spiking Neuron Models*. Cambridge University Press. [112, 128, 134]

Giese, M.: 1998, *Dynamic Neural Field Theory for Motion Perception*. Springer. [171, 185]

Giese, M. and T. Poggio: 2003, 'Neural mechanisms for the recognition of biological movements and actions'. *Nature Reviews Neuroscience* **4**, 179–192. [vi, 66, 67, 103, 104, 105, 106, 108, 169, 170, 182, 183]

Gollisch, T. and M. Meister: 2008, 'Rapid Neural Coding in the Retina with Relative Spike Latencies'. *Science* **319**, 1108–1111. DOI: 10.1126/science.1149639. [127, 163, 175]

Goncalves, L., E. DiBernardo, E. Ursella, and P. Perona: 1995, 'Monocular tracking of the human arm in 3D'. In: *Proceedings of the 5th International Conference on Computer Vision*. pp. 764–770. [101]

Goodale, M. A. and A. D. Milner: 1992, 'Separate visual pathways for perception and action'. *Trends in neurosciences* **15**(1), 20–25. [23]

Gorelick, L., M. Blank, E. Shechtman, M. Irani, and R. Basri: 2007, 'Actions as Space-Time Shapes'. *pami* **29**(12), 2247–2253. [101]

Grammont, F. and A. Riehle: 2003, 'Spike synchronization and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time'. *Biological cybernetics* **88**(5), 260–373. [128, 167, 180]

Graziano, M., R. Andersen, and R. Snowden: 1994, 'Tuning of MST neurons to spiral motions'. *The Journal of Neuroscience* **14**(1), 54–67. [42]

Grossberg, S. and E. Mingolla: 1985, 'Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading'. *Psychological review* **92**(2), 173–211. [71]

Grossberg, S., E. Mingolla, and L. Viswanathan: 2001, 'Neural dynamics of motion integration and segmentation within and across apertures'. *Vision Research* **41**(19), 2521–2553. [44, 71, 72, 73, 76]

Grossman, E., M.Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake: 2000, 'Brain Areas Involved in Perception of Biological Motion'. *Journal of Cognitive Neuroscience* **12**(5), 711–720. [104]

Grzywacz, N. and A. Yuille: 1990, 'A model for the estimate of local image velocity by cells on the visual cortex'. *Proc R Soc Lond B Biol Sci.* **239**(1295), 129–161. [48, 58, 59, 60, 61, 63, 80, 114, 164, 177]

Guichard, F. and L. Rudin: 1996, 'Accurate estimation of discontinuous optical flow by minimizing divergence related functionals'. In: *Proceedings of the International Conference on Image Processing*, Vol. I. pp. 497–500. [47]

Gupta, S. and J. Prince: 1996, 'On div-curl regularization for motion estimation in 3-d volumetric imaging'. In: *Proceedings of the International Conference on Image Processing*. pp. 929–932. [47]

Heeger, D. J.: 1992, 'Normalization of cell responses in cat striate cortex'. *Visual Neuroscience* **9**, 181–197. [57]

Hérault, J. and B. Durette: 2007, 'Modeling Visual Perception for Image Processing'. In: F. Sandoval, A. Prieto, J. Cabestany, and M. Gra na (eds.): *Computational and Ambient Intelligence : 9th International Work-Conference on Artificial Neural Networks, IWANN 2007*. [167, 179]

Hirai, M. and K. Hiraki: 2006, 'Neural Dynamics for Biological Motion Perception'. In: F. J. Chen (ed.): *Trends in Brain Mapping Research*. Nova Sciences Publishers, pp. 85–116. [3, 12, 104]

Hiris, E., D. Humphrey, and A. Stout: 2005, 'Temporal Properties in Masking Biological Motion'. *Perception and Psychophysics* **67**(3), 435–443. [169, 182]

Hogg, D.: 1983, 'Model-based vision: a paradigm to see a walking person'. *Image and Vision Computing* **1**(1), 5–20. [101]

Horn, B. and B. Schunck: 1981, 'Determining Optical Flow'. *Artificial Intelligence* **17**, 185–203. [47]

Huang, X., T. D. Albright, and G. R. Stoner: 2007, 'Adaptive Surround Modulation in Cortical Area MT'. *Neuron* **53**, 761–770. [4, 13, 37, 39, 92]

Huang, X., T. D. Albright, and G. R. Stoner: 2008, 'Stimulus Dependency and Mechanisms of Surround Modulation in Cortical Area MT'. *Journal of Neuroscience* **28**(51), 13889–13906. [37, 39]

Hubel, D. and T. Wiesel: 1962, 'Receptive fields, binocular interaction and functional architecture in the cat visual cortex.'. *J Physiol* **160**, 106–154. [23, 24, 25, 48]

Hubel, D. H. and T. N. Wiesel: 1960, 'Receptive fields of optic nerve fibres in the spider monkey'. *J. Physiol.* **154**, 572–80. [24]

Ilg, U.: 2008, 'The role of areas MT and MST in coding of visual motion underlying the execution of smooth pursuit'. *Vision Research* **48**(20), 2062–2069. [41]

Inaba, N., S. Shinomoto, S. Yamane, A. Takemura, and K. Kawano: 2007, 'MST neurons code for visual motion in space independent of pursuit eye movements'. *Journal of Neurophysiology* **97**(5), 3473–3483. [41]

Irani, M. and S. Peleg: 1993, 'Motion analysis for image enhancement: resolution, occlusion, and transparency'. *Journal on Visual Communications and Image Representation* **4**(4), 324–335. [47]

Izhikevich, E.: 2004, 'Which model to use for cortical spiking neurons?'. *IEEE Trans Neural Netw* **15**(5), 1063–1070. [131]

Jhuang, H., T. Serre, L. Wolf, and T. Poggio: 2007, 'A biologically inspired system for action recognition'. In: *Proceedings of the 11th International Conference on Computer Vision*. pp. 1–8. [vi, ix, 102, 103, 106, 107, 108, 110, 119, 120, 123, 137, 139, 168, 169, 170, 181, 183]

Jiang, H. and D. R. Martin: 2008, 'Finding Actions Using Shape Flows'. In: *Proceedings of the 10th European Conference on Computer Vision*, Vol. 5303 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pp. 278–292. [101]

Johansson, G.: 1973, 'Visual perception of biological motion and a model for its analysis'. *Perception and Psychophysics* **14**, 201–211. [3, 11, 103]

Jones, H., K. Grieve, W. Wang, and A. Sillito: 2001, 'Surround Suppression in Primate V1'. *Journal of Neurophysiology* **86**, 2011–2028. [28, 29, 60, 151, 164, 176]

Kara, P. and R. C. Reid: 2003, 'Efficacy of Retinal Spikes in Driving Cortical Responses'. *The Journal of Neuroscience* **23**(24), 8547–8557. [23, 24]

Kawano, K., M. Shidara, Y. Watanabe, and S. Yamane: 1994, 'Neural activity in cortical area MST of alert monkey during occular following responses'. *The Journal of Neurophysiology* **71**(6), 2305–2324. [42]

Kim, T.-K., S.-F. Wong, and R. Cipolla: 2007, 'Tensor Canonical Correlation Analysis for Action Classification'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Kreuz, T., J. S. Haas, A. Morelli, H. D. Abarbanel, and A. Politi: 2007, 'Measuring spike train synchrony'. *Journal of Neuroscience Methods* **165**, 151–161. [130, 163, 176]

Lagae, L., S. Raiguel, and G. A. Orban: 1993, 'Speed and direction selectivity of macaque middle temporal neurons'. *Journal of Neurophysiology* **69**(1), 19–39. [32, 34, 165, 178]

Landy, M. S. and J. R. Bergen: 1991, 'Texture Segregation and Orientation Gradient'. *Vision Research* **31**(4), 679–691. [69]

Laptev, I., B. Caputo, C. Schuldt, and T. Linderberg: 2007, 'Local velocity-adapted motion events for spatio-temporal recognition'. *Computer vision and image understanding* **108**, 207–229. [101, 102, 103]

Laptev, I., M. Marszalek, C. Schmid, and B. Rozenfeld: 2008, 'Learning realistic human actions from movies'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Laptev, I. and P. Perez: 2007, 'Retrieving actions in movies'. In: *Proceedings of the 11th International Conference on Computer Vision*. pp. 1–8. [102]

Levitt, J. and J. Lund: 1997, 'Contrast dependence of contextual effects in primate visual cortex'. *Nature* **387**(6628), 73–76. [30, 31]

Li, B., J. Thompson, T. Duong, M. Peterson, and R. Freeman: 2006, 'Origins of Cross-Orientation Suppression in the Visual Cortex'. *Journal of Neurophysiology* **96**, 1755–1764. [28]

Li, C. and W. Li: 1994, 'Extensive integration field beyond the classical receptive field of cat's striate cortical neurons–classification and tuning properties'. *Vision Research* **34**(18), 2337–2355. [29]

Liu, J., S. Ali, and M. Shah: 2008, 'Recognizing human actions using multiple features'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Liu, J. and W. T. Newsome: 2003, 'Functional Organization of Speed Tuned Neurons in Visual Area MT'. *Journal of Neurophysiology* **89**, 246–256. [114]

Liu, J. and M. Shah: 2008, 'Learning human actions via information maximization'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Livingstone, M. S. and B. R. Conway: 2003, 'Substructure of direction–selectivity receptive fields in macaque V1'. *Journal of Neurophysiology* **89**, 2743–2759. [24]

Lowe, D.: 2004, 'Distinctive image features from scale-invariant keypoints'. *International Journal of Computer Vision* **60**(2), 91–110. [102]

Lui, L. L., J. A. Bourne, and M. G. P. Rosa: 2007, 'Spatial Summation, End Inhibition and Side Inhibition in the Middle Temporal Visual Area MT'. *Journal of Neurophysiology* **97**(2), 1135. [36]

Majaj, N., M. Carandini, and M. J.A.: 2007, 'Motion Integration by Neurons in Macaque MT Is Local, Not Global'. *The Journal of Neuroscience* **27**(2), 366–370. [40]

Maldonado, P., C. Babul, W. Singer, E. Rodriguez, D. Berger, and S. Grün: 2008, 'Synchronization of Neuronal Responses in Primarly Visual Cortex of Monkeys Viewing Natural Images'. *Journal of Neurophysiology* **100**, 1523–1532. [167, 180]

Mante, V. and M. Carandini: 2005, 'Mapping of stimulus energy in primary visual cortex'. *Journal of Neurophysiology* **94**, 788—798. [80, 94]

Masson, G. and E. Castet: 2002, 'Parallel motion processing for the initiation of short-latency ocular following in humans'. *The journal of neuroscience* **22**(12), 5147–5163. [57]

Maunsell, J. and D. V. Essen: 1983, 'Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity.'. *J Neurophysiol* **49**, 1148–1167. [34]

Maunsell, J. H. and D. C. Van Essen: 1983, 'The Connections of the Middle Temporal Visual Area (MT) and their Relationship to a Cortical Hierarchy in the Macaque Monkey'. *The Journal of Neuroscience* **3**(12), 2563–2586. [32, 41]

Mestre, D. R., G. S. Masson, and L. S. Stone: 2001, 'Spatial scale of motion segmentation from speed cues'. *Vision Research* **41**(21), 2697–2713. [32]

Michels, L., M. Lappe, and L. Vaina: 2005, 'Visual areas involved in the perception of human movement from dynamic analysis'. *Brain Imaging* **16**(10), 1037–1041. [3, 12, 104]

Mikami, A., W. Newsome, and R. Wurtz: 1986, 'Motion selectivity in macaque visual cortex. II. Spatiotemporal range of directional interactions in MT and V1'. *Journal of Neurophysiology* **55**(6), 1328–1339. [34]

Milner, A. D. and M. A. Goodale: 2008, 'Two visual systems re–viewed'. *Neuropsychologia* **46**, 774–785. [23]

Mitra, S. and T. Acharya: 2007, 'Gesture Recognition: A Survey'. *IEEE Transactions on Systems, Man, and Cybernetics (SMC) – Part C: Applications and Reviews* **37**(3), 311–324. [101]

Moeslund, T. B., A. Hilton, and V. Kr"uger: 2006, 'A survey of advances in vision-based human motion capture and analysis'. *Computer Vision and Image Understanding* **104**(2–3), 90–126. [101]

Mokhber, A., C. Achard, and M. Milgram: 2008, 'Recognition of human behavior by space-time silhouette characterization'. *Pattern Recognition Letters* **29**(1), 81–89. [101]

Morrone, M., D. Burr, and L. Maffei: 1982, 'Functional implications of cross-orientation inhibition of cortical visual cells. I. Neurophysiological evidence'. *Proceedings of the Royal Society of London, Series B* **216**(1204), 335–354. [28]

Movshon, J., E. Adelson, M. Gizzi, and W. Newsome: 1986, 'The analysis of moving visual patterns'. *Experimental Brain Research* **11**, 117–151. [39]

Movshon, J., I. Thompson, and D. Tolhurst: 1978, 'Receptive field organization of complex cells in the cat's striate cortex'. *The Journal of Physiology* **283**(79–99). [25]

Movshon, J. A. and W. T. Newsome: 1996, 'Visual Response Properties of Striate Cortical Neurons Projecting to Area MT in Macaque Monkeys'. *Journal of Neuroscience* **16**(23), 7733–7741. [31]

Mutch, J. and D. G. Lowe: 2006, 'Multiclass Object Recognition with Sparse, Localized Features'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 11–18. [103, 106]

Nagel, H.: 1983, 'Constraints for the Estimation of Displacement Vector Fields from Image Sequences'. In: *International Joint Conference on Artificial Intelligence*. pp. 156–160. [47]

Nagel, H.: 1989, 'On a constraint equation for the estimation of displacement rates in image sequences'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 13–30. [47]

Nagel, H.-H.: 1987, 'On the estimation of optical flow: relations between different approaches and some new results'. *Artificial Intelligence Journal* **33**, 299–324. [47]

Nassi, J. J. and E. M. Callaway: 2006, 'Multiple Circuits Relaying Primate Parallel Visual Pathways to the Middle Temporal Area'. *Journal of Neuroscience* **26**(49), 12789–12798. [31]

Nassi, J. J. and E. M. Callaway: 2007, 'Specialized Circuits from Primary Visual Cortex to V2 and area MT'. *Neuron* **55**, 799–808. [32]

Nési, P.: 1993, 'Variational approach to optical flow estimation managing discontinuities'. *Image and Vision Computing* **11**(7), 419–439. [47]

Neuenschwander, S., M. Castelo-Branco, and W. Singer: 1999, 'Synchronous oscillations in the cat retina'. *Vision Research* **39**(15), 2485–2497. [128]

Niebles, J. and L. Fei-Fei: 2007, 'A Hierarchical Model of Shape and Appearance for Human Action Classification'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Niebles, J. C., H. Wang, and L. Fei-Fei: 2006, 'Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words'. In: *British Machine Vision Conference*. [103]

Niebles, J.-C., H. Wang, and L. Fei-Fei: 2008, 'Unsupervised Learning of Human Action Categories Using Spatial–Temporal Words'. *Internation Journal of Computer Vision* **79**(3), 299–318. [101]

Nowak, L. and J. Bullier: 1997, *The Timing of Information Transfer in the Visual System*, Vol. 12 of *Cerebral Cortex*, Chapt. 5, pp. 205–241. Plenum Press, New York. [127]

Nowlan, S. and T. Sejnowski: 1995, 'A selection model for motion processing in area MT of primates'. *J. Neuroscience* **15**, 1195–1214. [44, 60, 64]

Nowlan, S. J. and T. J. Sejnowski: 1994, 'Filter selection model for motion segmentation and velocity integration'. *j. Opt. Soc. Am. A* **11**(12), 3177–3199. [vi, 44, 60, 62, 64]

Odobez, J. and P. Bouthemy: 1995, 'Robust multiresolution estimation of parametric motion models'. *Journal of Visual Communication and Image Representation* **6**(4), 348–365. [47]

Ogata, T., W. Christmas, J. Kittler, and S. Ishikawa: 2006, 'Improving human activity detection by combining multi-dimensional motion descriptors with boosting'. In: *Proceedings of the International Conference on Pattern Recognition*, Vol. 1. Kowloon Tong, Hong Kong, pp. 295–298, comp-soc-press. [101]

Orban, G., F. Van Calenbergh, B. De Bruyn, and H. Maes: 1985, 'Velocity discrimination in central and peripheral visual field'. *Journal of the optical society of america, A* **2**(11), 1836–1847. [34]

Otte, M. and H. Nagel: 1994, 'Optical Flow Estimation: Advances and Comparisons'. In: J.-O. Eklundh (ed.): *Proceedings of the 3rd European Conference on Computer Vision*, Vol. 800 of *Lecture Notes in Computer Science*. pp. 51–70, Springer–Verlag. [46]

Pack, C. and R. Born: 2001, 'Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain'. *Nature* **409**, 1040–1042. [38, 60, 148, 150, 151]

Pack, C., B. Conway, R. Born, and M. Livingstone: 2006, 'Spatiotemporal Structure of Nonlinear Subunits in Macaque Visual Cortex'. *Journal of Neuroscience* **26**(3), 893–907. [25, 32]

Pack, C., A. Gartland, and R. Born: 2004, 'Integration of Contour and Terminator Signals in Visual Area MT of Alert Macaque'. *The Journal of Neuroscience* **24**(13), 3268–3280. [vii, 4, 13, 38, 148, 150, 151, 155, 157, 170, 184]

Pack, C. C., J. N. Hunter, and R. T. Born: 2005, 'Contrast Dependence of Suppressive Influences in Cortical Area MT of Alert Macaque'. *Journal of Neurophysiology* **93**(3), 1809–1815. [31]

Pack, C. C., M. S. Livingstone, K. R. Duffy, and R. T. Born: 2003, 'End-Stopping and the Aperture Problem: Two-Dimensional Motion Signals in Macaque V1'. *Neuron* **39**(4), 671–680. [38, 151]

Perkel, D. H. and T. H. Bullock: 1968, 'Neural coding'. *Neurosciences Research Program Bulletin* **6**, 221–348. [127]

Perrone, J.: 2004, 'A visual motion sensor based on the properties of V1 and MT neurons'. *Vision Research* **44**, 1733–1755. [44, 54, 114, 164, 177]

Perrone, J. and R. Krauzlis: 2008, 'Spatial integration by MT pattern neurons: a closer look at pattern-to-component effects and the role of speed tuning'. *Journal of Vision* **8**(9), 1–14. [40]

Perrone, J. and A. Thiele: 2001, 'Speed skills: measuring the visual speed analyzing properties of primate MT neurons'. *Nature Neuroscience* **4**(5), 526–532. [34, 54, 55, 114, 164, 177]

Pinto, N., D. D. Cox, and J. J. DiCarlo: 2008, 'Why is Real-World Visual Object Recognition Hard?'. *PLoS Comput Biol* **4**(1), e27. [168, 181]

Polana, R. and R. Nelson: 1997, 'Detection and recognition of periodic, non-rigid motion'. *ijcv* **23**(3), 261–282. [102]

Poppe, R.: 2007, 'Vision–based human motion analysis: An overview'. *Computer Vision and Image Understanding* **108**(1–2), 4–18. [101]

Potters, M. and W. Bialek: 1994, 'Statistical mechanics and visual signal processing'. *Journal de Physique I France* **4**, 1755–1775. [53]

Priebe, N., C. Cassanello, and S. Lisberger: 2003, 'The neural representation of speed in macaque area MT/V5'. *Journal of Neuroscience* **23**(13), 5650–5661. [28, 34, 35, 40, 114, 164, 177]

Priebe, N. J., S. G. Lisberger, and A. J. Movshon: 2006, 'Tuning for Spatiotemporal Frequency and Speed in Directionally Selective Neurons of Macaque Striate Cortex'. *The Journal of Neuroscience* **26**(11), 2941–2950. [25, 34, 164, 177]

Pucel, A. and D. Perret: 2003, 'Electrophysiology and brain imaging of biological motion'. *Philosophical transactions of the Royal Society B* **358**(1431), 435–445. [3, 12, 104]

Ragheb, H. and E. Hancock: 2003, 'A probabilistic framework for specular Shape-from-Shading'. *Pattern Recog.* **36**, 407–427. [101]

Reichardt, W.: 1957, 'Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems'. *Zeitschrift fur Naturforschung* **12**, 447–457. [52]

Riehle, A., S. Grün, M. Diesmann, and A. Aertsen: 1997, 'Spike Synchronization and Rate Modulation Differentially Involved in Motor Cortical Function'. *Science* **278**, 1950–1953. [167, 180]

Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek: 1997, *Spikes: Exploring the Neural Code*. Bradford Books. [128, 130]

Ringach, D. L.: 2002, 'Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex'. *Journal of Neurophysiology* **88**, 455–463. [25, 27, 48]

Rochel, O.: 2004, 'Une approche événementielle pour la modélisation et la simulation de neurones impulsionnels'. Ph.D. thesis, Université Henri Poincaré - Nancy 1. [172, 185]

Rodman, H. R. and T. D. Albright: 1989, 'Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT)'. *Experimental Brain Research* **75**, 53–64. [40]

Roelfsema, P. R., V. A. F. Lamme, and H. Spekreijse: 2004, 'Synchrony and covariation of firing rates in the primary visual cortex during contour grouping'. *Nature Neuroscience* **7**(9), 982–991. [128]

Rohr, K.: 1994, 'Toward model-based recognition of human movements in image sequences'. *CVGIP, Image Understanding* **1**, 94–115. [101]

Rust, N., V. Mante, E. Simoncelli, and J. Movshon: 2006, 'How MT cells analyze the motion of visual patterns'. *Nature Neuroscience* **9**, 1421–1431. [170, 184]

Saito, H., M. Yukie, K. Tanaka, K. Hikosaka, Y. Fukada, and E. Iwai: 1986, 'Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey'. *The Journal of Neuroscience* **6**(1), 145–157. [41]

Saul, A., P. Carras, and A. Humphrey: 2005, 'Temporal Properties of Inputs to Direction-Selective Neurons in Monkey V1'. *Journal of Neurophysiology* **94**, 282–294. [24, 81]

Sceniak, M., M. Hawken, and R. Shapley: 2001, 'Visual Spatial Characterization of Macaque V1 Neurons'. *Journal of Neurophysiology* **85**, 1873–1887. [29, 30, 60, 151, 164, 176]

Sceniak, M. P., M. J. Hawken, and R. Shapley: 2002, 'Contrast–dependent changes in spatial frequency tuning of macaque V1 neurons: effects of a changing receptive field size'. *Journal of Neurophysiology* **88**, 1363–1373. [31]

Sceniak, M. P., D. L. Ringach, M. J. Hawken, and R. Shapley: 1999, 'Contrast's effect on spatial summation by macaque V1 neurons'. *Nature Neuroscience* **2**(8), 733–739. [31]

Schindler, K. and L. J. Van Gool: 2008, 'Action snippets: How many frames does human action recognition require?'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–8. [102]

Schnörr, C.: 1991, 'Determining optical flow for irregular domains by minimizing quadratic functionals of a certain class'. *The International Journal of Computer Vision* **6**(1), 25–38. [47]

Schwabe, L., K. Obermayer, A. Angelucci, and P. Bressloff: 2006, 'The role of feedback in shaping the extra–classical receptive field of cortical neurons: a recurrent network model'. *The Journal of Neuroscience* **26**(36), 9117–9129. [30]

Seitz, S. and C. Dyer: 1997, 'View-invariant analysis of cyclic motion'. *The International Journal of Computer Vision* **25**(3), 231–251. [102]

Series, P.: 2002, "Etude th'eorique des modulations center/pourtour des propri'et'es des champs r'ecepteurs du cortex visuel primaire: circuits, dynamiques, et corr'elats perceptifs'. Ph.D. thesis, Universite de Paris-VI. [29]

Series, P., J. Lorenceau, and Y. Fregnac: 2003, 'The silent surround of V1 receptive fields: theory and experiments'. *Journal of physiology, Paris* **97**(4–6), 453–474. [29]

Serre, T.: 2006, 'Learning a dictionary of shape-components in visual cortex: Comparison with neurons, humans and machines'. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA. [103]

Serre, T., L. Wolf, and T. Poggio: 2005, 'Object recognition with features inspired by visual cortex'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 994–1000. [103, 106, 169, 183]

Shah, M. and R. Jain: 1997, *Motion-based recognition*, Computational Imaging and Vision Series. Kluwer Academic Publisher. [102]

Sigala, R., T. Serre, T. Poggio, and M. Giese: 2005, 'Learning Features of Intermediate Complexity for the Recognition of Biological Motion'. *ICANN 2005, LNCS 3696* pp. 241–246. [103, 106]

Simoncelli, E. P.: 1993, 'Dristributed Representation and Analysis of Visual Motion'. Ph.D. thesis, MIT Media Laboratory. [45]

Simoncelli, E. P. and D. Heeger: 1998, 'A Model of Neuronal Responses in Visual Area MT'. *Vision Research* **38**, 743–761. [vi, 44, 48, 64, 65, 66, 76, 81, 114, 164, 165, 170, 177]

Smith, A. T. and G. K. Edgar: 1990, 'The influence of spatial frequency on perceived temporal frequency and perceived speed'. *Vision Research* **30**(10), 1467–1474. [34]

Smith, M., N. Majaj, and A. Movshon: 2005, 'Dynamics of motion signaling by neurons in macaque area MT'. *Nature Neuroscience* **8**(2), 220–228. [39, 40]

Smith, M. A.: 2006, 'Surround Suppression in the Early Visual System'. *The Journal of Neuroscience* **26**(14), 3624–3625. [30]

Snowden, R. J., S. Treue, R. G. Erickson, and R. A. Andersen: 1991, 'The response of area MT and V1 neurons to transparent motion'. *The Journal of Neuroscience* **11**(9), 2768–2785. [32, 82, 165, 178]

Stoner, G. R. and T. D. Albright: 1992, 'Neural correlates of perceptual motion coherence'. *Nature* **358**, 412–414. [40]

Stumpf, P.: 1911, 'Über die Abhangigkeit der visuellen Bewegungsrichtung und negativen Nachbildes von den Reizvorgangen auf der Netzhaut'. *Zeitschrift fur Psychologie* **59**, 321–330. [149]

Suter, D.: 1994, 'Motion Estimation and Vector Splines'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Seattle, WA, pp. 939–942, IEEE. [47]

Tanaka, H. and I. Ohzawa: 2009, 'Surround Suppression of V1 Neuron Mediates Orientation–Based Representation of High–Order Visual Features'. *Journal of Neurophysiology* **101**, 1444–1462. [29]

Tanaka, K., K. Hikosaka, H. Saito, M. Yukie, Y. Fukada, and E. Iwai: 1986, 'Analysis of local and wide–field movements in the superior temporal visual areas of the macaque monkey'. *The Journal of Neuroscience* **6**(1), 134–144. [41]

Tanaka, K. and H. Saito: 1989, 'Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey'. *The Journal of Neurophysiology* **62**(3), 626–641. [41]

Thorpe, S.: 1990, 'Spike arrival times: A highly efficient coding scheme for neural networks'. *Parallel processing in neural systems and computers* pp. 91–94. [127]

Thorpe, S.: 2002, 'Ultra-Rapid Scene Categorization with a Wave of Spikes'. In: *Biologically Motivated Computer Vision*, Vol. 2525 of *Lecture Notes in Computer Science*. pp. 1–15, Springer-Verlag Heidelberg. [128]

Thorpe, S., A. Delorme, and R. VanRullen: 2001, 'Spike based strategies for rapid processing.'. *Neural Networks* **14**, 715–726. [163, 169, 175, 182]

Thorpe, S. and M. Fabre-Thorpe: 2001, 'Seeking categories in the brain'. *Science* **291**, 260–263. [169, 182]

Thorpe, S., D. Fize, and C. Marlot: 1996, 'Speed of processing in the human visual system'. *Nature* **381**, 520–522. [127]

Thurau, C. and V. Hlavac: 2008, 'Pose primitive based human action recognitio in videos or still images'. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*. pp. 1–6. [101]

Tistarelli, M.: 1995, 'Computation of coherent optical flow by using multiple constraints'. In: *Proceedings of the 5th International Conference on Computer Vision*. pp. 263–268. [46]

Tlapale, É., G. S. Masson, and P. Kornprobst: 2008, 'Motion Integration Modulated by Form Information'. In: *Deuxième conférence française de Neurosciences Computationnelles*. [4, 13, 76, 169, 171, 183, 184]

Todorovic, D.: 1996, 'A gem from the past: Pleikart Stumpf's (1911) anticipation of the aperture problem, Reichardt detectors, and perceived motion loss at equiluminance'. *Perception* **25**(10), 1234–1242. [149]

Topsoe, F.: 2000, 'Some Inequalities for Information Divergence and Related Measures of Discrimination'. *IEEE Transactions on information theory* **46**(4), 1602–1609. [115]

Tran, D. and A. Sorokin: 2008, 'Human activity recognition with metric learning'. In: *Proceedings of the 10th European Conference on Computer Vision*, Vol. 5302 of *LNCS*. pp. 548–561, Springer–Verlag. [101]

Ungerleider, L. and M. Mishkin: 1982, *Two cortical visual systems.*, pp. 549–586. MIT Press. [23]

Vaina, L., J. Solomon, S. Chowdhury, P. Sinha, and J. Belliveau: 2001, 'Functional neuroanatomy of biological motion perception in humans'. *Proceedings of the National Academy of Science* **98**(20), 11656–11661. [104]

Van Essen, D. C. and J. L. Gallant: 1994, 'Neural mechanisms of form and motion processing in the primate visual system'. *Neuron* **13**, 1–10. [23]

Van Santen, J. and G. Sperling: 1984, 'Temporal covariance model of human motion perception'. *Journal of the Optical Society of America A* **1**(5), 451–473. [52]

Van Santen, J. and G. Sperling: 1985, 'Elaborated Reichardt detectors'. *Journal of the Optical Society of America A* **2**(2), 300–320. [48, 52, 53]

VanRullen, R. and S. J. Thorpe: 2002, 'Surfing a spike wave down the ventral stream'. *Vision Research* **42**, 2593–2615. [127, 163, 175]

Victor, J. and K. Purpura: 1996, 'Nature and precision of temporal coding in visual cortex: a metric-space analysis.'. *J Neurophysiol* **76**, 1310–1326. [128, 130]

Walker, G. A., I. Ohzawa, and R. D. Freeman: 1999, 'Asymmetric Suppression Outside the Classical Receptive Field of the Visual Cortex'. *The Journal of Neuroscience* **19**(23), 10536–10553. [28, 29, 30, 31]

Wallach, H.: 1935, 'Über visuell wahrgenommene Bewegungsrichtung'. *Psychological Research* **20**(1), 325–380. [150]

Wang, D. L. and D. Terman: 1995, 'Locally excitatory globally inhibitory oscillator networks'. *IEEE Trans. Neural Net.* **6**, 283–286. [128]

Wang, L. and D. Suter: 2007, 'Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model'. In: *Proceedings CVPR*. [101]

Watson, A. and A. Ahumada: 1983, 'A look at motion in the frequency domain'. *NASA Tech. Memo.* [48, 50, 52, 53, 85, 136]

Watson, A. and A. Ahumada: 1985, 'Model of human visual-motion sensing'. *J Opt Soc Am A* **2**(2), 322–342. [48, 50, 56, 85]

Webb, B. S., N. T. Dhruv, S. G. Solomon, C. Tailby, and P. Lennie: 2005, 'Early and Late Mechanisms of Surround Suppression in Striate Cortex of Macaque'. *The Journal of Neuroscience* **25**(50), 11666–11675. [29, 30, 31]

Wielaard, D. J., M. Shelley, D. McLaughlin, and R. Shapley: 2001, 'How Simple Cells Are Made in a Nonlinear Network Model of the Visual Cortex'. *The Journal of Neuroscience* **21**(14), 5203–5211. [131]

Wilson, H., V. Ferrera, and C. Yo: 1992, 'A psychophysically motivated model for two-dimensional motion perception.'. *Visual Neuroscience* **9**(1), 79–97. [vi, 69, 70, 171, 184]

Wohrer, A. and P. Kornprobst: 2009, 'Virtual Retina : A biological retina model and simulator, with contrast gain control'. *Journal of Computational Neuroscience* **26**(2), 219. DOI 10.1007/s10827-008-0108-4. [128, 167, 179]

Wong, S.-F. and R. Cipolla: 2007, 'Extracting Spatiotemporal Interest Points using Global Information'. In: *Proceedings of the 11th International Conference on Computer Vision*. pp. 1–8. [102]

Wong, S.-F., T.-K. Kim, and R. Cipolla: 2006, 'Learning Motion Categories using both Semantic and Structural Information'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–6. [102]

Wong, S.-F., T.-K. Kim, and R. Cipolla: 2007, 'Learning Motion Categories Using Both Semantic and Structural Information'. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. pp. 1–6. [101, 103]

Wuerger, S., R. Shapley, and N. Rubin: 1996, '"On the visually perceved direction of motion" by Hans Wallach: 60 years later'. *Perception* **25**, 1317–1367. [150]

Xiao, D., S. Raiguel, V. Marcar, J. Koenderink, and G. A. Orban: 1995, 'Spatial Heterogeneity of Inhibitory Surrounds in the Middle Temporal Visual Area'. *Proceedings of the National Academy of Sciences* **92**(24), 11303–11306. [93, 113, 133, 166, 170, 178, 179, 183]

Xiao, D.-K., V. Marcar, S. Raiguel, and O. G.A.: 1997a, 'Selectivity of Macaque MT/V5 Neurons for Surface Orientation in Depth Specified by Motion'. *European Journal of Neuroscience* **9**, 956–964. [36]

Xiao, D. K., S. Raiguel, V. Marcar, and G. A. Orban: 1997b, 'The spatial distribution of the antagonistic surround of MT/V5 neurons.'. *Cereb Cortex* **7**(7), 662–77. [36, 39, 92, 93, 113, 133, 166, 170, 179, 183]

Yilmaz, A. and M. Shah: 2008, 'A differential geometric approach to representing the human actions'. *Computer vision and image understanding* **119**(3), 335–351. [101]

Yo, C. and H. Wilson: 1992, 'Perceived direction of moving two-dimensional patterns depends on duration, contrast and eccentricity.'. *Vision Res* **32**(1), 135–47. [57]

Zaksas, D. and T. Pasternak: 2005, 'Area MT Neurons Respond to Visual Motion Distant From Their Receptive Field'. *Journal of Neurophysiology* **94**, 4156–4167. [32]

Zelnik-Manor, L. and M. Irani: 2001, 'Event-based analysis of video'. In: *Proceedings of CVPR'01*, Vol. 2. pp. 123–128. [101, 103]

Zelnik-Manor, L. and M. Irani: 2006, 'Statistical Analysis of Dynamic Actions'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(9), 1530–1535. [115]

Zhu, G., G. W. Xu, Changsheng, and Q. Huang: 2006, 'Action recognition in broadcast tennis video using optical flow and support vector machine'. In: *Proceedings of the 9th European Conference on Computer Vision*, Vol. 3979 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pp. 89–98. [101]