Distributed neural computation for the visual perception of motion

Mauricio Cerda

Nancy-Université
Université Nancy 2

UFR mathématiques et informatique **École doctorale IAEM Lorraine**

Département de formation doctorale en informatique

Calcul neuronal distribué pour la perception visuelle du mouvement

THÈSE

présentée et soutenue publiquement le 14 Octobre 2011

pour l'obtention du

Doctorat de l'université Nancy 2

(spécialité informatique)

par

Mauricio David Cerda Villablanca

Composition du jury

Président :	Le président	
Rapporteurs :	Mathias QUOY	Professeur, Université de Cergy-Pontoise, France
	Adrian PALACIOS	Professeur, Universidad de Valparaiso, Chili
Examinateurs :	Heiko NEUMANN	Professeur, University of Ulm, Allemagne
	Anne BOYER	Professeur, Université Nancy 2, France
	Rachid DERICHE	Directeur de Recherche, INRIA, Sophia-Antipolis, France
	Bernard GIRAU (directeur)	Professeur, Université Henri Poincaré, Nancy 1

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

Résumé

Le travail présenté dans cette thèse propose des modèles de calcul pour l'extraction du mouvement et la reconnaissance de formes dynamiques à partir du flux d'informations visuelles, en s'inspirant des mécanismes correspondants mis en jeu dans le cerveau. Plus précisément, nous proposons des hypothèses sur la façon dont le mécanisme cérébral de ces tâches peut fonctionner et nous nous efforçons de déterminer comment des neurones avec un petit champ récepteur sont en mesure de fournir des réponses cohérentes et de coder des formes dynamiques complexes. Nous étudions chaque aspect du traitement réalisé dans le cerveau que nous avons modélisé dans un cadre connexionniste, en montrant comment ces systèmes distribués peuvent être utilisés pour des tâches complexes telles que la détection de mouvement et la reconnaissance de formes dynamiques. Du point de vue informatique ces modèles offrent de nouveaux algorithmes, avec des propriétés intéressantes telles que l'utilisation de mémoire distribuée et la robustesse.

La détection de mouvement et la discrimination de motifs visuels complexes à partir de ce signal (ou "vision cognitive") structurent les deux parties dans lesquelles le manuscrit se divise. La première partie porte sur la détection de mouvement en étudiant la façon dont l'extraction de caractéristiques visuelles est effectuée à partir du flux d'information visuel, et en particulier la façon dont les problèmes dus à la petite taille et la gamme de détection réduite des détecteurs de mouvement locaux peuvent être résolus. Dans la deuxième partie nous étudions la façon dont la classification des motifs visuels dynamiques complexes est réalisée à partir du traitement fourni par le système primaire de vision pour réaliser ce que nous appelons la vision cognitive, en évaluant au passage différentes techniques d'extraction de caractéristiques visuelles.

Mots-clés: Sciences cognitives, perception du mouvement, mouvement biologique, réseaux neuronaux.

Abstract

The work presented in this thesis proposes computational models for motion extraction and pattern recognition from the visual flow of information in the brain and determine how the two tasks can be understood together. More precisely, we propose hypotheses about how the brain mechanism for these tasks may work and we seek to show how neurons with a small receptive field are able to deliver coherent answers and encode complex patterns. We study each aspect of brain processing that we have modelled in a connectionist framework, showing how such distributed systems can be used in complex tasks such as motion detection and pattern recognition. From the computer science perspective, these models provide new algorithms, with interesting properties such as distributed memory utilization and robustness.

We have focused our work on two aspects of visual information processing: the detection of motion and the discrimination of complex visual patterns from that signal (or "cognitive vision") that constitute the two parts in which this work is divided. The first part of this thesis about motion detection studies how feature extraction is performed from the visual flow of information, specifically how problems due to the small size and range of the local motion detectors can be solved. In the second part, we work on how the classification of complex visual patterns is achieved from the processing provided by the early vision system, evaluating different feature-extraction techniques to perform what we call cognitive vision.

Keywords: Cognitive sciences, motion perception, biological motion, neural networks.

Dedicada a mi esposa Daniela...

Contents

Introduction

Life sciences have searched to understand how the brain works, involving an increasing number of fields, including Computer Science. Meanwhile Computer Science has often taken inspiration from biology to propose new (and better) ways of computation. It is in this context, that two separated fields have come together in what is called computational neuroscience, a field where biology and computer science meet with the objective of understanding how the brain processes information at both general and detailed levels. Computer science explores biology in search of inspiration and new computational methods, while biologists search for new insights that modern mathematical and computational tools can provide to their work. To give an example, the action potential propagation model proposed by Hodgkin-Huxley in 1952 provides a precise theoretical description of the electrical propagation of activity in a neuron (they worked with the squid giant axon), yet it is still very difficult to study theoretically a network of such neurons. Even for much simpler models of neurons, the analysis that can be made is limited. However, computer simulations can provide great insights into how a network of neurons can behave; currently existing neural network simulators are capable of taking into account even the geometry of each neuron. As another example, a well-known technique to perform data classification is the Multilayer Perceptron proposed by Rumelhart et al in 1986, a model directly inspired by the idea of neurons and layers of units. Of course this is not a model of how real neurons work but the principle of having many units computing something together comes from the understanding of how biological networks work.

For more than sixty years, through the works of Hubel & Wiesel on the cat brain, we have known that neurons in the brain receive visual information from the retina, with each neuron processing only a small part of the visual field. These neurons are located in the most external layer of the brain: the cortex, measuring 2 to 4 mm in depth (in humans). This is in itself surprising, as we perceive the visual world as one, and not divided into small pieces. To date, the functions of many neurons in the cortex have been determined, especially in the primary visual cortex, which has been the subject of numerous studies. However, the principle of a local input for each unit remains in the primary visual cortex. But, how are populations of these units in general able to deliver and encode responses with local information? This question has motivated a large number of recent works in neuroscience. The visual treatment of information in the brain has commonly been associated with two flows of information since the works of Ungerleider and Mishkin in 1982: first, the ventral pathway dealing with static information, and second, a dorsal pathway commonly associated with movement. The two, take as input different parallel channels going from the retina and through the thalamus with an increasing size of receptive field along paths in the visual cortex. Despite the large number of neurons discovered with different response profiles, it remains an open question how complex patterns are encoded if features are extracted locally. To give an example, until the 1990's the mainstream theory about how faces where stored proposed a hierarchical decomposition into a set a of features. However, a recent work by Freiwald and colleagues indicates that pairs and even trios of features are encoded by single neurons, refuting the hypothesis of a hierarchy and supporting the idea of decomposition

into sets of pairs/trios of features for each face. These examples raise two questions: (1) how are local detectors able to provide coherent responses? And (2) how can they encode complex patterns as a population?

This thesis proposes computational models for extracting motion and pattern recognition from the visual flow of information in the brain and determining how the two tasks can be understood together. In this work we propose a hypothesis as to how the actual brain mechanism for these tasks works and we seek to answering how local detectors are able to deliver coherent answers and encode complex patterns. In each aspect of brain processing that we have studied and modelled, we analyze the problem in a connectionist framework showing how such distributed systems can be used in complex tasks such as pattern recognition. From the computer science perspective these models provide new algorithms with interesting properties such as distributed memory utilization and robustness to units failures. As we mention below, we have focused our work on two aspects of visual information processing: early vision and the discrimination of complex visual patterns, or what we call "cognitive vision". In the first part of this thesis about early vision, we study how feature extraction is performed from the visual flow of information: how light is perceived and movement is extracted. In the second part we work on how local patterns are detected and used for the discrimination of visual patterns, or how the classification of complex visual patterns is achieved from the processing provided by the early vision system, to perform what we call cognitive vision.

Early vision

The early vision system in mammals deals with the extraction of local features, *i.e.* features from the information of a small part of the visual field, where the result of this processing is not specific to some cognitive task. There is a wide variety of feature detectors that have been observed in mammals and other species such as: movement, orientation, color, disparity or combinations, such as movement near corners and movements at different depths. The element that links these detectors is their spatially localized action. However, local information is not always sufficient to provide coherent answers. In such cases, considering both the local treatment of information and the dynamic of the system (propagation of information) provides the right feature extraction.

In this part we focus on two fundamental problems of vision linked to the propagation of information: how local detectors can handle the aperture problem, *i.e.* how to propagate coherent information, and the way a set of local movement detectors can provide a wide detection of speeds, in other words, how detectors tuned to different ranges of speed provide a global population response. For the aperture problem, we propose a two-layer mechanism with a feedback architecture to make the detection converge to *the less ambiguous signal* and to provide a local detection where the aperture problem is solved. Here we show that the feedback architecture of two layers with different receptive field sizes can drive the system to propagate the less ambiguous signal. The range of speed problem studies how to perform speed detection in a wider range of speed than the range of any particular detector. In this part we show that this can be satisfactorily obtained if each detector output is considered altogether with a confidence measure for its detection, and uniformly distributed in the log speed space.

The two models, one for motion detection, and the other for the range of speed detection, show how the combination of local computation and the dynamics of a population can solve two fundamental problems when motion detection is performed locally. For both models we perform simulations with artificial sequences to validate our results.

Cognitive Vision

To recognize subjects, to remember the name of a song, to identify a smell are examples of cognitive tasks, but what characterizes them? In order to carry out these tasks we require an initial feature extraction: some information about the external world. However, feature extraction is not the only element that cognitive tasks require for processing information. There is also an important element of experience of the subject that drives this processing. Considering these elements, local feature extraction and previous experiences, we can wonder how does the brain process them together?

In order to study how visual patterns are encoded and extracted in the brain, we must first verify if local optical flow patterns (with larger receptive fields than in motion extraction) can be statistically constructed from local features, similar to the response of the neurons of some primates or to experimental results in psychophysics. This kind of operator is biologically plausible and we show statistically that it is a significant way to encode optical flow for synthetic and realistic video sequences. We conclude that pattern recognition may use such operators (and not local information directly) to have both: pseudo scale-invariant detectors and a distributed representation. Given the local (or pseudo-local) nature of the features we deal with, how can a cognitive task such as visual sequence discrimination be achieved? Here we propose again that it is the dynamics of the population that provides this capacity, presenting a coding strategy where each complex pattern is decomposed into a set of local spatiotemporal trajectories.

Our model proposes, from a computational point of view, a way to encode patterns as sets of local features, keeping the local nature of the processing, only requiring an observer to perform the population readout, reinforcing the idea of distributed processing of information even for cognitive tasks such as visual pattern discrimination. From a biological point of view, we propose a possible mechanism to account for the pattern encoding and recognition. This mechanism can explain some of the observed properties from the literature, such as rapid temporal discrimination, support of time compression/dilatation, viewer rotation variance and point-light stimuli recognition.

The main contributions of this thesis are the proposed models for motion detection and pattern classification. These two models were proposed in a distributed framework, showing how the combined responses within a population of units can deliver the right respond, even if each unit receives information only from close units. We study early vision tasks in the first part, where we propose a distributed population of units that can successfully deliver a coherent answer (see Chapter 2) and a larger range of detection that of any particular detector (see Chapter 3). In the second part, (see Chapters 4 and 5) we present a distributed model to perform pattern recognition, showing how local features can be used to solve this type of task. Using the model, we show that the use of local optical flow patterns as feature, being both statistically significant and biologically plausible (see Chapter 6), improves the overall discrimination performance. Our model presents a hypothesis about how visual pattern recognition in the brain may be performed, not by saving the moving/static templates but by recording the local features dynamics.

Résumé étendu

Le travail présenté dans cette thèse propose des modèles de calcul pour l'extraction du mouvement et la reconnaissance de formes dynamiques à partir du flux d'informations visuelles, en s'inspirant des mécanismes correspondants mis en jeu dans le cerveau. Ces deux tâches (extraction et reconnaissance) sont étudiées également afin de mieux comprendre comment elles peuvent interagir ensemble. Plus précisément, nous proposons des hypothèses sur la façon dont le mécanisme cérébral de ces tâches peut fonctionner et nous nous efforçons de déterminer comment des neurones avec un petit champ récepteur sont en mesure de fournir des réponses cohérentes et de coder des formes dynamiques complexes. Nous étudions chaque aspect du traitement réalisé dans le cerveau que nous avons modélisé dans un cadre connexionniste, en montrant comment ces systèmes distribués peuvent être utilisés pour des tâches complexes telles que la détection de mouvement et la reconnaissance de formes dynamiques. Du point de vue informatique ces modèles offrent de nouveaux algorithmes, avec des propriétés intéressantes telles que l'utilisation de mémoire distribuée et la robustesse.

Comme nous l'avons mentionné, nous avons concentré notre travail sur deux aspects du traitement de l'information visuelle : la détection de mouvement et la discrimination de motifs visuels complexes à partir de ce signal, que nous désignons par " vision cognitive". Ces deux aspects structurent les deux parties dans lesquelles le manuscrit se divise. La première partie de cette thèse, constituée des trois premiers chapitres, porte sur la détection de mouvement en étudiant la façon dont l'extraction de caractéristiques visuelles est effectuée à partir du flux d'information visuel, et en particulier la façon dont les problèmes dus à la petite taille et la gamme de détection réduite des détecteurs de mouvement locaux peuvent être résolus. Dans la deuxième partie, constituée des trois derniers chapitres, nous étudions la façon dont la classification des motifs visuels dynamiques complexes est réalisée à partir du traitement fourni par le système primaire de vision pour réaliser ce que nous appelons la vision cognitive, en évaluant au passage différentes techniques d'extraction de caractéristiques visuelles.

Dans la suite nous allons décrire les différents chapitres qui forment ce manuscrit de thèse, en commençant par une brève description de chacun d'eux, avant de décrire plus en profondeur l'ensemble du travail.

1. **Chapitre 1, Vision et mouvement.**
   Il s'agit d'un chapitre d'introduction et d'état de l'art où la détection locale du mouvement est présentée selon deux points de vue : l'approche de la vision par ordinateur et l'approche biologique. Cette étude montre que, indépendamment du substrat (ordinateurs ou neurones), la détection locale du mouvement pose des problèmes fondamentaux communs en raison de la nature locale du traitement. Parmi ces différents problèmes, nous choisissons d'étudier le problème d'ouverture et le problème de l'échantillonnage de la vitesse. L'intérêt est d'étudier comment ces problèmes peuvent être résolus par le cerveau des primates, de façon à proposer des algorithmes bio-inspirés pouvant résoudre ces problèmes, tout en discutant différentes hypothèses en biologie sur les mécanismes spécifiques mis en œuvre pour résoudre ces problèmes.

2. **Chapitre 2, Le problème d'ouverture.**
   Ce chapitre présente un modèle distribué de résolution du problème d'ouverture, avec un état de l'art plus précis en lien avec ce problème bien connu de la perception visuelle du mouvement. Dans le modèle présenté dans ce chapitre, une inhibition latérale est introduite afin d'améliorer la résolution du problème d'ouverture lorsque la détection locale de mouvement est bruitée. Des comparaisons avec un modèle établi, où aucune inhibition n'entre en jeu, sont effectuées pour des séquences d'images synthétiques et réelles.
3. **Chapitre 3, Echantillonnage de la vitesse.**
   Ce chapitre étudie le lien entre l'erreur relative dans l'estimation de la vitesse et les capacités de discrimination de la vitesse observées chez l'humain. Suite à cette comparaison, nous présentons un modèle de discrimination de la vitesse qui propose de déterminer la façon de choisir et de combiner des détecteurs locaux de vitesse de manière à obtenir une discrimination de vitesse dans une gamme et avec une précision similaires à celles rencontrées chez l'homme. Des comparaisons de notre modèle avec des résultats biologiques sont présentées sur des séquences synthétiques.
4. **Chapitre 4, Caractéristiques visuelles et discrimination.**
   Ce chapitre présente un état de l'art sur l'extraction de caractéristiques visuelles et la discrimination de séquences temporelles de signaux visuels, en particulier liés au mouvement humain, selon les points de vue de la vision par ordinateur et de la biologie. Cette étude montre que le codage de séquences en 2D est biologiquement plus plausible, éventuellement en utilisant les informations de mouvement local et un modèle simple du mouvement humain. En termes de caractéristiques visuelles extraites, nous montrons que l'information locale est importante lorsque des séquences de mouvement humain sont prises en considération. Nous soulignons également dans ce chapitre que plusieurs caractéristiques visuelles peuvent être considérés pour discriminer les séquences visuelles, mais que les aspects temporels et les relations spatiales sont connus pour être essentiels dans l'analyse clinique du mouvement et la psychophysique, bien qu'ils soient souvent négligés en vision par ordinateur.
5. **Chapitre 5, Discrimination de séquences.**
   Dans ce chapitre nous explicitons l'idée que les séquences visuelles temporelles de mouvement humain peuvent être décomposées en un ensemble de trajectoires spatio-temporelles locales, comme énoncé d'abord dans le chapitre 4. Nous expliquons comment une population d'unités (les neurones) peut de façon répartie coder une séquence visuelle. Le modèle présenté tient compte de la position et de la vitesse de chaque trajectoire, où nous développons une stratégie analytique pour déterminer les paramètres de chaque unité de la population de manière à coder une séquence visuelle donnée. Pour évaluer le modèle présenté, une série d'expériences est effectuée sur des séquences synthétiques pour vérifier que la population d'unités représente bien la position et la vitesse des trajectoires voulues une fois que nous avons configuré de manière appropriée les paramètres.
6. **Chapitre 6, Evaluation.**
   Ce chapitre évalue le modèle présenté au chapitre 5, dans le cadre de séquences de mouvement humain réelles. Tout d'abord la propriété de discrimination est vérifiée exclusivement en termes de trajectoires (orientation, vitesse), à partir de trajectoires directement fournies pour des parties spécifiques du corps (articulations). Après ces expériences, nous vérifions si le flux optique peut être utilisé directement comme signal d'entrée dans notre modèle sans fournir explicitement les trajectoires des parties du corps. Dans la deuxième partie

du chapitre, une sélection statistique de caractéristiques visuelles locales est introduite et évaluée dans le contexte de notre modèle de discrimination.

7. **Conclusions.**
   Pour conclure au sujet de notre travail, ce chapitre reprend et présente les principales contributions de la thèse, ainsi qu'une discussion sur les perspectives futures.
8. **Références.**
9. **Annexes.**
   Certaines parties techniques et mathématiques citées au cours des différents chapitres sont regroupées en annexe.

I Traitement primaire de la vision

Le système de vision primaire des mammifères permet l'extraction des caractéristiques visuelles locales, *i.e.* les caractéristiques issues de l'information d'une petite partie du champ visuel, où le résultat de ce traitement n'est pas spécifique à une tâche cognitive donnée. Il existe une grande variété de détecteurs de caractéristiques visuelles qui ont été observés chez les mammifères et chez d'autres espèces, tels que pour: le mouvement, l'orientation, la couleur, la disparité ou des combinaisons comme le mouvement des coins, les mouvements à différentes profondeurs, etc. L'élément qui relie ces détecteurs est leur caractère local. Toutefois, les informations locales seules ne sont pas toujours suffisantes pour apporter des réponses cohérentes, par exemple dans le cas de la détection de mouvement. Dans la première partie de cette thèse, nous mettons l'accent sur la détection locale du mouvement, principalement parce que le mouvement est une caractéristique robuste, qui peut être déjà obtenue en niveaux de gris, dans des conditions monoculaires et même en présence de bruit. Cette détection est basée sur les changements temporels dans le flux d'information visuelle, généralement associés à des aires où l'information visuelle est riche. Parce que la détection de mouvement est confrontée à de nombreux écueils, nous nous concentrons sur deux problèmes que nous jugeons importants : le problème d'ouverture et le problème d'échantillonnage de la vitesse. Pris ensemble, ces problèmes portent sur l'orientation et l'amplitude du mouvement, deux aspects importants de l'estimation locale du mouvement.

Le problème d'ouverture

Le problème d'ouverture restreint l'estimation du mouvement à la seule composante de mouvement qui est perpendiculaire au contour local, quand on regarde la source visuelle à travers une petite ouverture. Le principe de faire appel à des calculs locaux sur les images pour en extraire le mouvement aide à fournir des implantations efficaces logicielles ou matérielles spécialisées dans le calcul du flux optique. Mais comme ce principe est local, il se heurte inévitablement au problème d'ouverture. Malgré les différences dans les schémas de codage de l'information visuelle, des neurones ont été trouvés dans le cerveau avec un profil de réponse qui peut être associé à un détecteur de vitesse au niveau local (ces neurones se situant dans les aires V1-MT). Cette similitude motive l'étude de la perception du mouvement chez les primates, pour comprendre comment le cerveau traite le problème fondamental d'ouverture, voir la figure 1. Le problème d'ouverture n'est bien entendu pas le seul problème qui se pose lorsqu'un mouvement est détecté au moyen de détecteurs locaux, mais c'est un bon exemple pour comparer les points de vue informatique et biologique.

Dans le cortex visuel des primates, les neurones de l'aire V1 ont un champ récepteur réduit, et la plupart d'entre eux sont sensibles à une certaine fréquence spatio-temporelle. Lorsque le

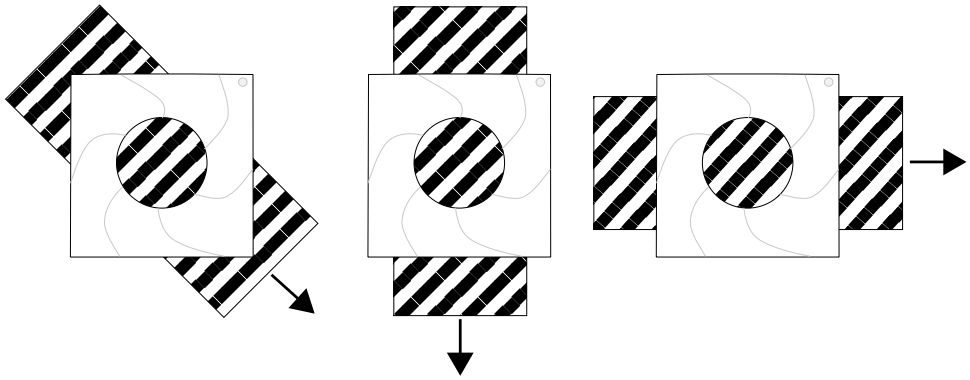


FIGURE 1 – Le problème d'ouverture. Au travers de l'ouverture, les trois mouvements apparaissent identiques (reproduit à partir de [AB85]).

mouvement est ambigu dans son champ récepteur, en raison du problème d'ouverture, l'activité de ces neurones n'est sensible qu'à une composante donnée du mouvement. Les neurones de V1 ne sont ainsi sensibles qu'à la seule composante de mouvement perpendiculaire au contour [MN96], et sont ainsi sujets au problème d'ouverture. D'autre part les unités dans l'aire MT, après un certain délai à partir de leur première activation (50 à 75 ms), sont activés par le mouvement réel, résolvant ainsi le problème d'ouverture. Le champ récepteur des neurones dans l'aire MT est plus grand que le champ récepteur des neurones de V1, par un facteur de 5 et une structure complexe a été notée à l'intérieur de ce champ récepteur, où des orientations cohérentes semblent renforcer une orientation donnée tandis que des orientations opposées semblent l'inhiber [LPB01], la forme exacte de cette interaction n'étant pas entièrement connue.

Compte tenu de ces éléments, nous proposons un mécanisme à deux couches avec une architecture de rétroaction pour faire converger la détection vers le signal le moins ambigu et pour fournir une détection locale où le problème d'ouverture est résolu, voir les figures 2(a) et 2(b). Notre principale contribution est de montrer que l'idée de propagation dynamique issue de [BN04] peut être améliorée par un mécanisme bio-inspiré de compétition qui permet de maintenir localement la prééminence des filtres les plus activés tout en renforçant les détecteurs qui correspondent à la vitesse réelle seule. Grâce à ce principe, une perception du mouvement cohérente se propage le long des contours de l'objet en mouvement, avec une meilleure tolérance au bruit.

Nous appliquons notre méthode à des séquences vidéo synthétiques et réelles, de façon à montrer qu'il converge de manière similaire à la méthode proposée par [BN04]. C'est dans les scénarios bruités que la différence entre les deux modèles apparaît. La figure 3 montre les résultats pour la séquence réelle. Dans ce cas, notre méthode montre une meilleure tolérance au bruit, avec des signaux parasites (en dehors du damier il n'y a pas de mouvement) qui sont éliminés après quelques itérations, tout en extrayant correctement le mouvement du damier. Conjointement avec une approche distribuée efficace pour effectuer la détection de mouvement en résolvant le problème d'ouverture, notre modèle propose une interprétation du microcircuit trouvé dans l'aire MT, où des effets de facilitation et d'inhibition ont été signalés pour les neurones de détection de la vitesse.

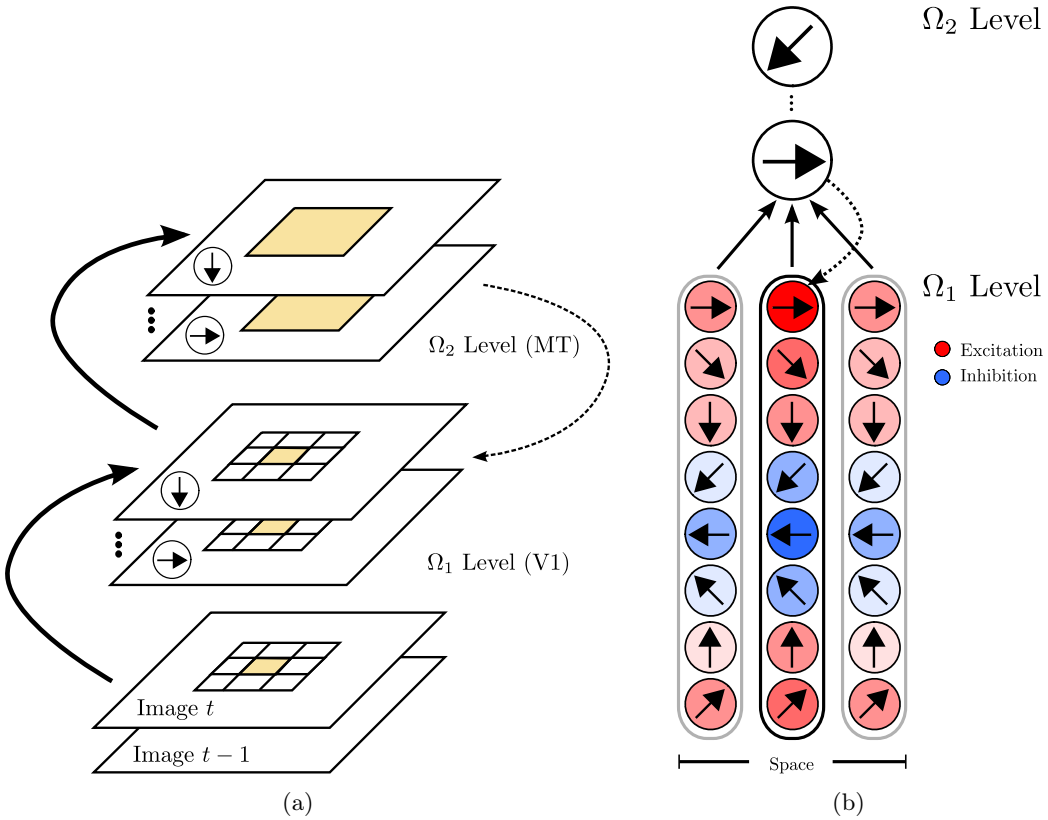


FIGURE 2 – Les lignes continues désignent les connexions directes (feed-forward) et la ligne pointillée une connexion de rétroaction (feedback). (a) L'architecture générale de notre modèle pour résoudre le problème d'ouverture. A chaque pixel sont associés plusieurs détecteurs de mouvement, autant que de vitesses à détecter. La seconde couche a un champ récepteur plus large que la première. (b) Détails sur le lien entre la première et la seconde couche. Chaque unité dans la deuxième couche reçoit une influence positive à partir des orientations similaires (en rouge) et une inhibition à partir des orientations opposées (en bleu). En même temps la distance spatiale module la force de l'interaction.

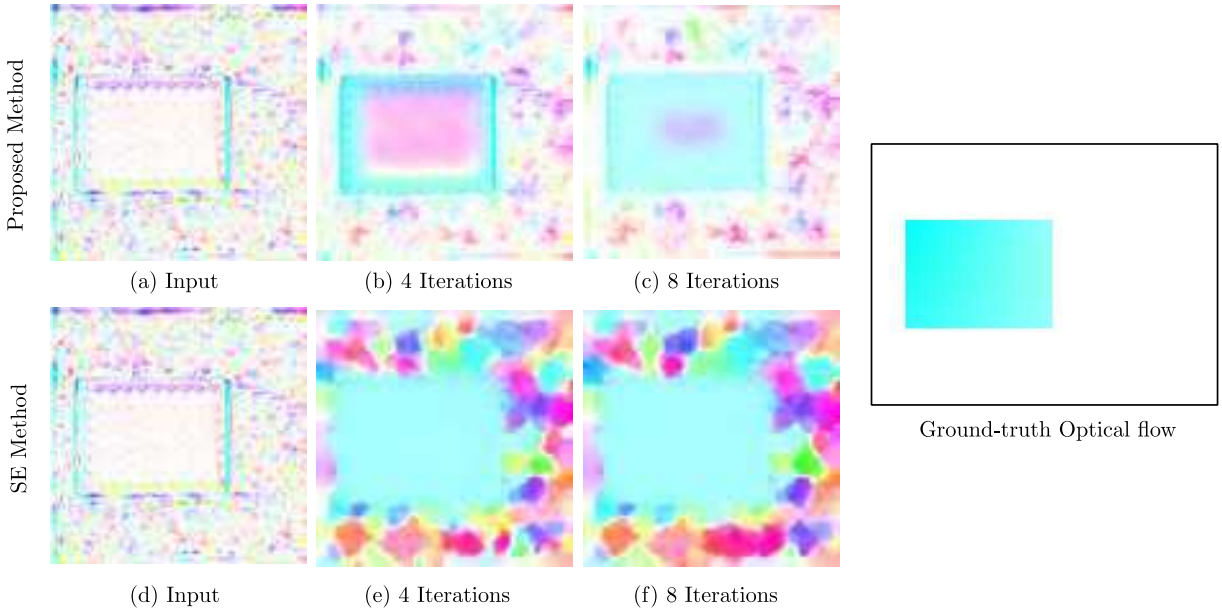


FIGURE 3 – L'évolution de la détection de vitesse pour un damier en mouvement de translation. Nous utilisons notre modèle AEI (première rangée) et à titre de comparaison le modèle SE (deuxième rangée). La dernière colonne indique le signal de mouvement véritable, et de gauche à droite les modèles à 0, 4 et 8 itérations.

Echantillonnage de la vitesse

Un autre problème avec la détection locale de motion, c'est qu'elle implique la détection d'une gamme limitée de vitesses. Ce problème se pose à la fois en vision par ordinateur et dans le cerveau des mammifères : comment une gamme de vitesse plus grande que la gamme d'un seul détecteur peut-elle être construite ? La vision par ordinateur propose généralement un mécanisme où le même signal doit être analysé à de multiples échelles, en commençant par la plus élevée, et en exécutant une série d'approximations, voir la figure 4(a). D'après différentes expériences menées, le cerveau semble avoir choisi une stratégie très différente en tenant compte en parallèle des différentes échelles (à la fois dans le temps et dans l'espace), et en combinant en quelque sorte ces estimations [MVB94, NHD05] (les vitesses plus ou moins élevées ne nécessitent pas plus de temps pour être estimées) de façon à obtenir un large éventail de détection de vitesse avec une erreur relative (et non pas absolue) d'estimation de la vitesse constante, voir la figure 4(b). La différence exacte entre les deux stratégies et le mécanisme neuronal précis qui permet d'obtenir une gamme étendue de détection motivent notre étude de l'échantillonnage de vitesses.

En considérant le principe d'un calcul parallèle, et afin d'éviter la propagation d'erreurs entre les différentes échelles, nous proposons d'avoir une mesure de confiance pour chaque détecteur de vitesse, en fonction de la vitesse détectée, de façon à tenir compte préférentiellement du détecteur le plus approprié en fonction de chaque vitesse. Afin de réaliser ceci, nous proposons une mesure de confiance pour chaque détecteur de vitesse en fonction de chaque échelle spatiale, et une moyenne pondérée pour combiner les réponses de l'ensemble des détecteurs de vitesse.

La méthode que nous proposons pour l'échantillonnage et la combinaison s'inspire de la physiologie humaine et de connaissances psychophysiques en ce sens qu'elle aboutit aux propriétés de discrimination relative uniforme sur une large plage de vitesses en utilisant un ensemble de

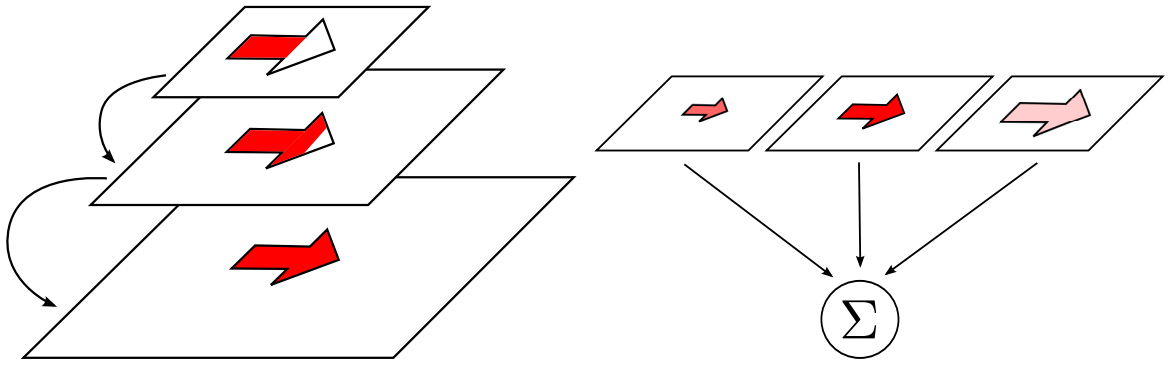


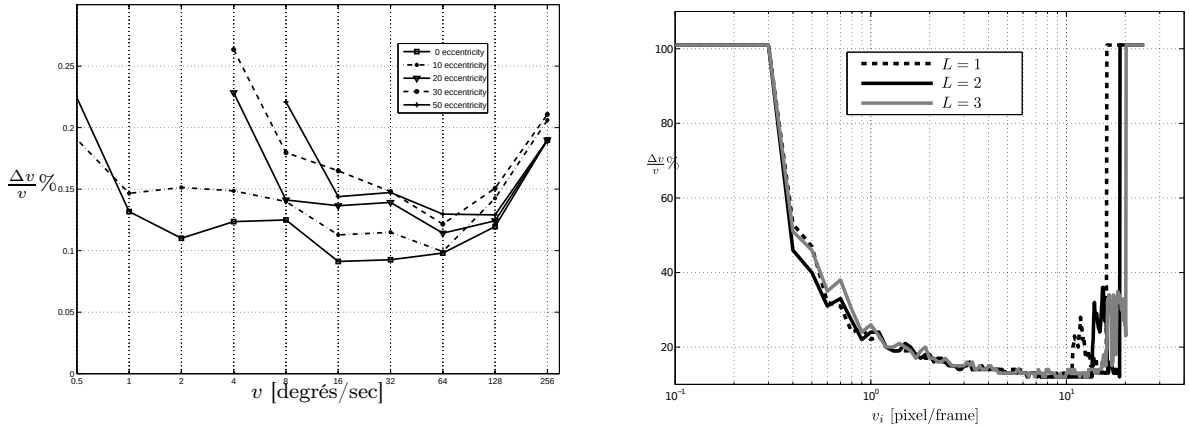
FIGURE 4 – Dans l'approche sérielle (a) la confiance, en rouge, augmente à mesure que estimation se propage. L'approche parallèle (b) estime simultanément plusieurs vitesses à différents niveaux de confiance.

détecteurs régulièrement distribués dans l'espace logarithmique, et qu'elle donne des résultats en temps constant en fonction des échelles, voir la figure 5. Ainsi, sur la base de cette idée, nous pouvons estimer le nombre de détecteurs de mouvement différents que nous devrions considérer (pour chaque orientation) pour une plage de vitesses donnée, dans laquelle la réponse émergera de l'interaction des différents détecteurs locaux de mouvement. Notre modèle peut aussi être compris comme un possible codage par population du mouvement, où, pour une vitesse donnée, plusieurs unités peuvent être activées (chacune avec une vitesse préférée). La méthode de combinaison des détecteurs de vitesse que nous proposons peut, au moins, vérifier la propriété d'homogénéité observée dans les expériences psychophysiques de discrimination de vitesses.

En termes de calcul distribué de mouvement, notre modèle explique comment des détecteurs de mouvements locaux peuvent être organisés de manière à avoir des capacités de discrimination de vitesse similaires à l'homme, avec une technique qui peut être encore comparable à des algorithmes de vision par ordinateur, en reliant l'erreur relative des détecteurs à la vitesse discriminée. De plus, le calcul parallèle de la vitesse que nous proposons ne dépend d'aucun processus sériel, même si une combinaison des réponses doit être effectuée sur la population complète. Prises ensembles, les solutions que nous proposons pour le problème d'ouverture et désormais pour le problème d'échantillonnage de la vitesse présentent une vue d'ensemble de la façon d'estimer le mouvement dans le contexte d'un calcul local, où deux problèmes importants ont été abordés en rapport avec l'orientation et l'amplitude du mouvement.

Bilan de la première partie

L'ouverture et l'analyse multi-échelle dans la détection du mouvement sont deux problèmes fondamentaux de la vision (sans être les seuls) observés dans les systèmes artificiels et dans le cerveau, malgré la très grande différence de substrat de calcul. Des solutions radicalement différentes semblent être proposées dans le cadre de chaque approche, et nous avons présenté deux modèles informatiques de la façon dont le cerveau pourrait résoudre ces problèmes dans le contexte d'un calcul totalement distribué. Nous montrons comment un système distribué peut résoudre des problèmes qui ne peuvent pas être résolus en utilisant uniquement des informations locales, et où la dynamique de la population conjuguée à l'information locale fournit la solution aux problèmes, en l'occurrence ici le problème d'ouverture et de la détection de vitesse multi-échelles. Dans le



(a) Capacité humaine en discrimination de vitesse : 5%

(b) Discrimination de vitesse réalisée par notre modèle : 20%.

FIGURE 5 – Capacité en discrimination de vitesses (a) de l'homme et (b) de la méthode proposée. Le pourcentage représente la quantité relative de variation de vitesse nécessaire pour distinguer les deux mouvements.

cas du problème d'ouverture, nous sommes en mesure de lever l'ambiguïté des réponses locales en réalisant un consensus local et en propageant cette information à travers l'espace et le temps. Pour l'échantillonnage de la vitesse, nous montrons qu'une répartition appropriée des détecteurs locaux de mouvement (que nous précisons) peut permettre une détection de vitesse avec une gamme donnée dans laquelle une erreur relative constante sera atteinte. Ainsi, cet aspect tend à montrer qu'une distribution appropriée de l'estimateur de mouvement local (dans l'espace des vitesses) détermine la plage globale et la précision de la population totale.

II Vision cognitive

Nous avons étudié dans la première partie de cette thèse le calcul de la vitesse, car il s'agit d'un élément important dans la compréhension des scènes visuelles. Comme nous l'avons vu, l'estimation du mouvement peut être vue comme un calcul distribué, mais cette approche nécessite du temps supplémentaire pour éliminer les ambiguïtés, par exemple pour résoudre le problème d'ouverture (environ 100 à 200 ms), voir la figure 6 pour les valeurs de référence du temps écoulé jusqu'à la première réponse dans différentes aires du cerveau impliquées dans la reconnaissance des séquences de mouvement complexes. En même temps, il y a une erreur intrinsèque due à la nature discrète de l'échantillonnage de la vitesse, qui est relative à la vitesse estimée (environ 5 %), donc à vitesse plus élevée l'erreur absolue est plus grande. En dépit de cela, des tâches telles que la reconnaissance de mouvements biologiques (marcher, courir, sauter) où l'information de mouvement est un élément crucial, prennent très peu de temps, fournissant par exemple la catégorisation au bout de 120 à 200 ms [PD07] selon le protocole. Cette tâche ne semble pas dépendre de la précision absolue de l'estimation locale de la vitesse. Par conséquent, ce genre de tâche cognitive ne peut pas dépendre de l'extraction de caractéristiques précises, mais d'aspects qui peuvent être calculés presque directement. Cela implique une tolérance aux erreurs que nous savons être intrinsèques à l'estimation locale du mouvement, comme le problème d'ouverture et les erreurs relatives commises dans l'estimation de la vitesse. Pour étudier ce codage

des séquences spatio-temporelles complexes, nous nous concentrons sur un type particulier de séquences, le mouvement biologique lié aux mouvements du corps et du visage, pour lequel de nombreuses expériences sont disponibles.

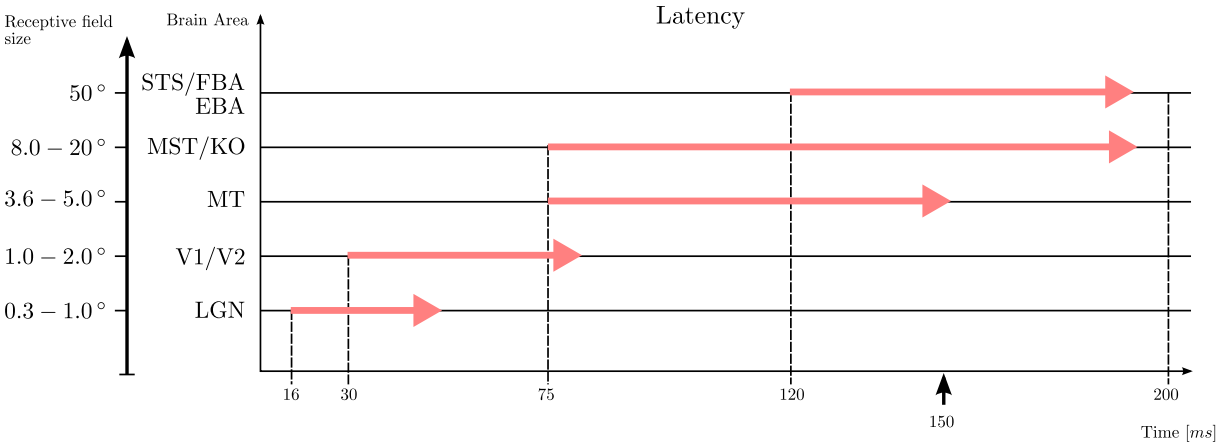


FIGURE 6 – Temps de latence des différentes aires du cerveau impliquées dans la reconnaissance de séquences visuelles (mouvement biologique). La latence est le temps requis pour la première activation de l'aire après que le stimulus visuel a été présenté. On notera que la taille de la flèche n'est pas à l'échelle. L'activité n'est pas instantanée, car elle débute et elle peut durer pendant plusieurs centaines de ms dans certaines aires du cerveau. Le diagramme montre également le champe récepteur (de référence) de chaque aire.

Une expérience particulière utilisant des stimuli par points lumineux (Point-Light PL), où seules les articulations d'un acteur sont éclairées, a fourni de nombreux indices sur le traitement du mouvement biologique. Différentes études utilisant les stimuli PL [TG08, CG05] indiquent que la caractéristique visuelle la plus pertinente pour effectuer la classification des séquences de mouvements de l'homme (au au moins pour les stimuli PL) est le mouvement local. Cela a été testé à l'aide de variantes des stimuli PL où et un intervalle inter-stimuli a été utilisé entre les images pour montrer que la reconnaissance est fortement diminuée si les caractéristiques locales sont perturbées [MRS92]. Pour des stimuli PL identiques, la phase relative de chaque point est cruciale pour identifier le même motif, et la reconnaissance humaine est extrêmement sensible aux variations de ce type [BP94] pour le mouvement biologique. En termes de vitesse précise, Shiffrar et al. [SLC97] ont montré que même si le mouvement biologique est vu au travers de plusieurs petites ouvertures, donc de manière ambiguë, il peut encore être reconnu. En outre, il a été montré que l'augmentation du nombre de PL facilite la reconnaissance [BP94], à partir d'au moins 8 à 10 marqueurs d'articulations.

En termes d'aires corticales impliquées dans la reconnaissance des formes complexes, des expériences menées par [SP09] à l'aide d'IRM fonctionnelle confirment qu'il y a au moins une aire responsable du mécanisme utilisé pour identifier les séquences visuelles complexes (aire appelée Sillon Temporal Supérieur ou STS, Superior Temporal Sulcus), en utilisant des informations de mouvement ou des informations statiques, ainsi que proposé par [GP03]. Cette aire présente une activation quand un motif visuel complexe est présenté, et elle présente une plus grande activation lorsque le même motif est en mouvement. Dans le même temps d'autres études ont montré que le codage des séquences visuelles complexes est plus probablement 2D et n'est ni invariant aux rotations dans le plan de l'observateur [PP03], ni aux perturbations 3D [BBS98]

(des stimuli PL).

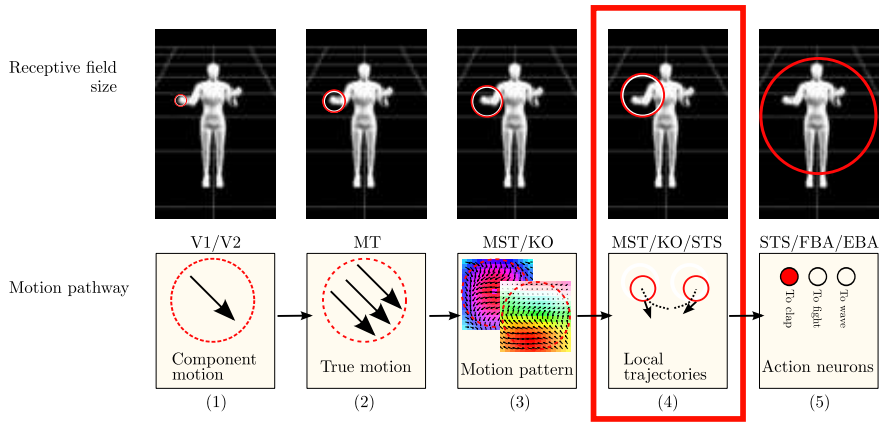


FIGURE 7 – L'architecture générale proposée dans notre modèle. Notre contribution principale est à l'étape 4, où nous représentons la séquence comme un ensemble de caractéristiques locales. La fonction de l'étape 3 sera examinée après la présentation de l'étape 4 de notre modèle.

Au total, ces idées peuvent être comprises comme un système de codage de séquences complexes, basé sur un ensemble de caractéristiques visuelles locales, où il est très important de garder une trace de la structure spatiale du motif au cours du temps. Une configuration statique pourrait activer le système, mais les mouvements devront l'activer encore plus, car il correspondra à une certaine prédiction de mouvement. En outre, la représentation devrait être 2D, montrant une légère invariance à la rotation dans le plan de l'observateur. L'entrée principale d'un tel mécanisme est la détection de mouvement située dans l'espace, pas nécessairement très précise en termes d'orientation ou de vitesse, mais précisément située dans l'espace associé aux mouvements codés, voir la figure 7 pour une illustration générale des différentes étapes de traitement dans le codage des séquences visuelles complexes. Dans la deuxième partie de cette thèse, nous développons un modèle considérant ces aspects, de façon à montrer que les aspects spatiaux du mouvement peuvent être utilisés pour coder des séquences complexes dans un cadre distribué qui s'inspire de la reconnaissance de mouvements biologiques chez l'homme.

Discrimination de séquences

Étant donné le caractère local des caractéristiques visuelles, comme nous l'avons expliqué, leur localisation dans l'espace peut rapidement fournir des informations pour identifier les séquences visuelles complexes. Nous proposons à nouveau une interprétation selon laquelle c'est la dynamique de la population qui fournit une telle discrimination, en présentant une stratégie de codage où chaque motif complexe est décomposé en un ensemble de trajectoires spatio-temporelles locales. Notre modèle propose, à partir d'un point de vue informatique, une manière de coder les motifs comme des ensembles de caractéristiques locales, en maintenant le caractère local du traitement tout en introduisant le besoin d'un observateur pour effectuer la combinaison de la population de détecteurs, ce qui renforce l'idée d'un traitement distribué de l'information, même pour des tâches cognitives telles que la la discrimination visuelle des formes. D'un point de vue biologique, nous proposons un mécanisme possible pour réaliser le codage et la reconnaissance des motifs, où les motifs globaux induisent des prévisions locales à comparer aux entrées fournies.

Puisque nous considérons que la séquence peut être réduite en plusieurs trajectoires perti-

nentes définies au cours du temps (selon l'idée sous-jacente aux stimuli PL), comme nous l'avons présenté, nous considérons que si nous sommes capables de coder ces trajectoires, nous pouvons coder les séquences. Cela implique que nous supposons que nous sommes en mesure de connaître la position de ces points dans le temps, et nous voulons différencier les trajectoires. Nous supposons ainsi dans notre modélisation qu'il est possible d'extraire des caractéristiques locales de façon à obtenir plusieurs trajectoires spatio-temporelles, idée que nous évaluons dans la section suivante, et nous nous concentrons sur la reconnaissance de motifs dynamiques dans les séquences. Afin de différencier les trajectoires spatiales nous étendons un modèle de calcul basé sur la théorie des champs neuronaux continus (CNFT) [Ama77, Tay99, XG02], où l'entrée visuelle est projetée sur une population 2D d'unités ou neurones. Nous commençons par une description de la théorie CNFT, un cadre mathématique permettant de modéliser les populations de neurones.

- **Théorie des champs neuronaux continus (Continuous neural field theory, CNFT).** Dans la CNFT, l'activité des neurones (potentiel de membrane) est représentée par leur fréquence de décharge moyenne, où les populations sont des cartes continues de neurones, sans délai de transmission entre les unités [Ama77] et avec une dépendance linéaire à l'intensité de l'entrée. Le potentiel de membrane m suit alors l'équation 1:

$$\frac{\partial m(x,t)}{\partial t} + \tau m(x,t) = \int_{\Omega} w(|x' - x|) f[m(x',t)] dx' + I(x,t) + h \tag{1}$$

  où I est l'activité d'entrée (entrée rétinienne dans le cas du cortex visuel, ou des informations de mouvement dans notre modèle de discrimination), H est le seuil de neurone, f est la fonction d'activation des neurones et l'intégration se fait sur l'ensemble des neurones Ω. Bien sûr, le cortex ressemble davantage à une variété 2D que 1D, mais pour des raisons de simplicité, nous fondons notre explication sur le cas 1D. Parmi les différents types de solutions, un type très étudié de solution pour m correspond aux différentes formes de bulles d'activité. Ces bulles peuvent être utilisées de manière à modéliser des motifs d'activité auto-soutenue, la propagation de fronts d'activité [Coo05] ou la compétition entre des populations [Tay99]. Dans ces différents cas il s'agit de formes d'activité qui ont été observées dans le cortex humain et d'autres structures du cerveau, comme décrit par Wu et al [WXC08]. Dans la figure , nous illustrons le motif de propagation de front, étant donnée une entrée se déplaçant à la vitesse v, l'activité en $m(x,t)$ forme un front d'activité qui se déplace à la même vitesse que l'entrée avec un certain retard. Dans cet exemple la fonction noyau $w(|x' - x|)$ est symétrique.
- **CNFT appliquée aux séquences.** De façon à coder des trajectoires spatio-temporelles, par exemple comme les stimuli PL, nous avons développé une version modifiée de la CNFT, où la fonction noyau w n'est pas symétrique, mais asymétrique, pour tenir compte de la vitesse sur la trajectoire, voir figure . Le principal résultat que nous établissons dans cette nouvelle version de la CNFT est que, comme l'activité est de plus en plus localisée spatialement, nous pouvons toujours approcher la relation entre l'asymétrie (β) et la vitesse d'entrée (v) par une fonction linéaire afin de maximiser l'activité totale. Ce résultat théorique nous permet de coder n'importe quelle trajectoire en 1D. Dans le même temps pour des cas précis d'entrée et de noyau (versions analytiques), nous pouvons obtenir cette relation de manière précise.

$$v \approx \frac{\beta}{2\tau} \tag{2}$$

  Pour vérifier ce résultat, nous vérifions le codage de plusieurs trajectoires synthétiques à des

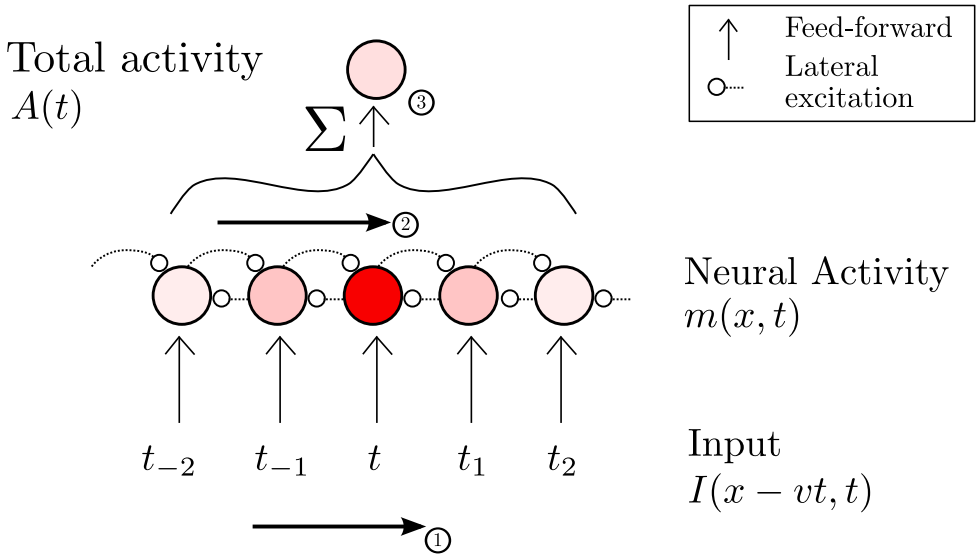


FIGURE 8 – L'architecture générale de la CNFT. Etant donnée une activité d'entrée se déplaçant à la vitesse v (1), le champ neuronal peut être réglé (via le profil du noyau $w(x'-x)$) pour avoir un front d'activité se déplaçant à la vitesse de l'entrée (2), mais l'activité totale sera impossible à distinguer de celle obtenue à partir d'une entrée se déplaçant à la vitesse $-v$ par exemple. Dans notre modèle, nous mesurons l'activité totale avec une seule unité (3). La couleur rouge représente l'activité neuronale. Dans la CNFT il y a aussi des connexions latérales inhibitrices de longue portée (non représentées), et elles sont symétriques.

vitesses différentes, pour montrer que notre modèle est capable de coder les deux propriétés simultanément. La principale contrainte que nous avons identifiée dans notre expérience est que tant que les trajectoires se chevauchent, la discrimination n'est pas possible, mais si elles diffèrent au moins partiellement, les séquences globales peuvent être différenciées. En outre, comme nous représentons le mouvement visuel en 2D, nous avons besoin de coder non seulement la vitesse, mais aussi l'orientation. Afin d'aborder cette question, nous avons étendu le modèle ACNFT en 2D, en considérant une rotation de la même fonction noyau asymétrique, orientée dans la direction du mouvement.

Le modèle CNFT asymétrique que nous avons introduit montre comment le codage de séquences spatio-temporelles pourrait être effectué dans le cerveau, non pas par un codage explicite des vues 2D successives de la scène visuelle (comme le principal modèle actuel le propose) mais en enregistrant implicitement les caractéristiques locales qui décrivent un modèle global, au travers des connexions synaptiques. Notre modèle ne dit pas nécessairement que la reconnaissance humaine de séquences visuelles repose entièrement sur des motifs locaux. Des expériences telles que les stimuli PL aléatoires (illumination aléatoire des parties du corps) ont d'ailleurs montré que cette interprétation ne bénéficie pas d'indices très solides. Cependant, nous soutenons l'idée qu'il est possible que les motifs locaux soient dérivés d'une représentation interne du corps pour les reconnaître, comme un "simulateur" interne (ou de certains indices de forme). Un bon candidat pour étayer cette hypothèse est fourni par les neurones miroirs [RFG01], pour lesquels une activation a été observée à chaque fois que nous voyons quelqu'un en mouvement. Le lien entre ces éléments (la discrimination et la cinématique interne) a été déjà exploré. Dans [JS04] l'exécution d'une tâche motrice a montré une influence sur la perception visuelle des mêmes tâches ou d'autres tâches, par exemple si l'observateur marche sa capacité à distinguer d'autres sujets qui

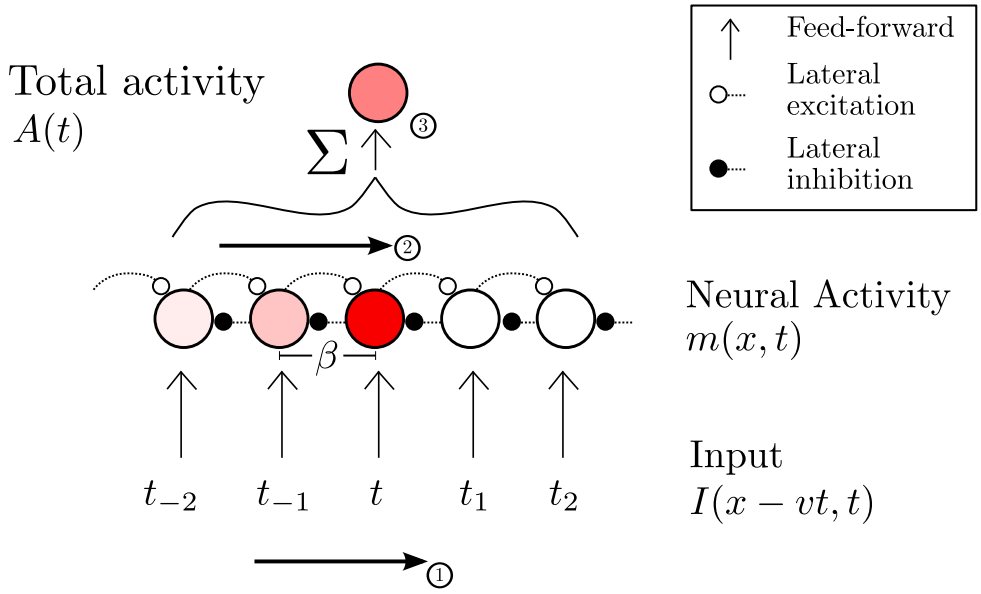
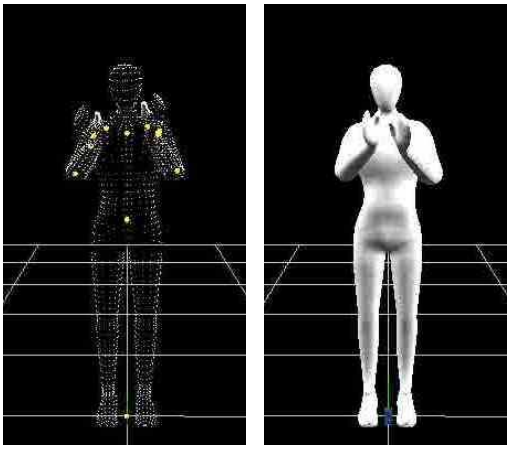


FIGURE 9 – Comportement de notre variante de la CNFT avec une forme asymétrique pour le noyau $w(x' - x - \beta)$ (ACNFT). Etant donnée une activité d'entrée se déplaçant à la vitesse v (1), le champ neuronal peut être réglé (via le profil du noyau $w(x' - x - \beta)$) de façon à avoir un front d'activité se déplaçant à une certaine vitesse donnée en entrée (2), de manière à avoir pour cette vitesse la plus haute activité totale dans (3). La couleur rouge représente l'activité neuronale. Dans le modèle ACNFT il y a également des connexions à longue distance, qui sont inhibitrices (non représentées), et asymétriques.



(a) (b)

FIGURE 10 – L'un des mouvements que nous considérons: "applaudir". (a) La version PL où seuls les articulations sont indiquées. (b) L'animation de la même séquence combinant les articulations, la maillage du corps et l'information de squelette.

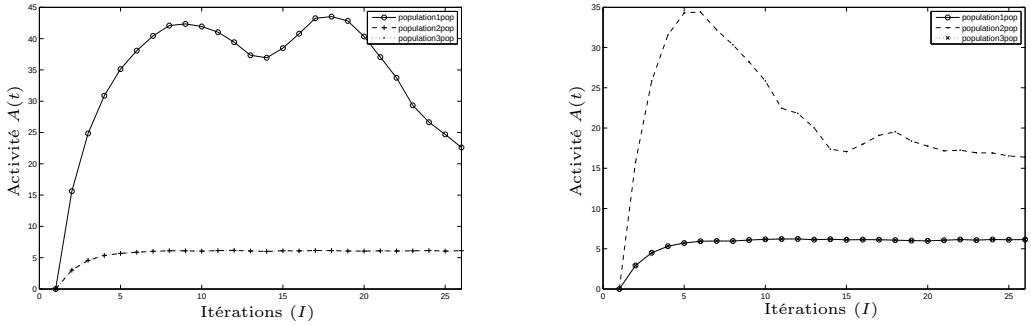
marchent va baisser. Cette interférence entre ce que nous voyons et ce que nous pouvons anticiper pourrait aussi expliquer la baisse de performance pour des stimuli PL issus d'animaux [PS09], dans la mesure où nous sommes plus familiers avec les mouvements humains qu'avec la cinématique animale.

Le modèle présenté illustre comment une population d'unités, le champ de neurones, peut coder un motif global par le biais d'une représentation distribuée. La seule contrainte est que le motif doit être décomposable en trajectoires locales. Notre modèle distribué pour la classification de séquences de mouvement illustre comment non seulement des tâches "simples" telles que la détection de mouvement peuvent être résolues par une approche distribuée, mais également des tâches de plus haut niveau telles que la classification des séquences temporelles.

Evaluation

Dans la dernière partie de notre travail, nous évaluons la performance du modèle ACNFT présenté à l'aide de trajectoires locales capturées à partir d'un ensemble de caméras, afin de vérifier notre modèle et d'étudier ses propriétés dans des conditions réalistes. Nous évaluons également la performance de notre modèle en utilisant deux méthodes possibles d'extraction des caractéristiques : (1) le flux optique brut et (2) la décomposition en motifs locaux de mouvement qui semble être suggérée comme une entrée appropriée à la discrimination de séquences, afin de vérifier l'éventuelle contribution de ce mécanisme. L'objectif principal est de confronter notre modèle proposé avec des séquences de mouvement réel simultanément capturé en 3D et en 2D, et de discuter quelles propriétés des mécanismes de reconnaissance de l'action humaine on peut expliquer.

Nous avons d'abord vérifié la discrimination réalisée par le modèle ACNFT en mesurant l'activité totale (dans l'espace et le temps) de 3 populations : "applaudir", "combattre", "saluer", une pour chaque motif considéré, et en utilisant comme entrée les trois trajectoires possibles du poignet (en 2D), voir figure 10. L'activité est $4,14$ fois, $11,3$ fois et $4,28$ fois plus élevée pour la population adéquate, ce qui signifie que lorsque l'entrée correspond à la séquence "applaudir",



(a) (b)

FIGURE 11 – (a) L'activité de la population m_1 ("applaudir") en fonction des itérations, pour les trois entrées possibles, une activité plus élevée correspond à l'entrée "applaudir". L'activité est un peu plus élevée pour la population "applaudir". (b) Idem pour la population "combattre".

l'activité totale pour cette population est au moins $4,14$ fois supérieure à celle des deux autres populations (dans le pire des cas). On peut également remarquer que l'activité pour les autres populations n'est pas nulle, car même si le noyau ne correspond à aucune trajectoire, il existe quand même une (plus faible) activation. A partir des valeurs observées, on remarque que la trajectoire de "applaudir" semble plus difficile à représenter, dans la mesure où l'écart absolu entre la mauvaise entrée et l'entrée codée est plus petit.

Une des propriétés que nous essayons de récupérer avec notre modèle est la classification rapide, donc nous regardons l'activité totale au fil du temps et nous répétons le même test avec les trois stimuli, comme dans la figure 11. Dans chaque test, l'activité moyenne de la population est sensiblement différente de celle des deux autres populations, en utilisant le test statistique One-Way ANOVA ($p < 0,05$) [Sap06], et les deux autres populations ne sont pas significativement différentes entre elles, en utilisant un test multiple entre les populations ($p < 0,05$). En outre, les deux courbes se chevauchent dans la figure 11 par exemple. On peut remarquer que l'activité temporelle diffère très rapidement, plus précisément à partir de la deuxième itération. Donc la discrimination peut être réalisée très rapidement. Cependant, l'activité n'est pas constante, parce que les mouvements que nous considérons n'ont pas des vitesses constantes. Indépendamment des différentes vitesses, le codage de l'ACNFT est efficace pour coder les mouvements même avec des vitesses irrégulières, permettant dans le même temps une discrimination très rapide.

Nous analysons également comment notre modèle réagit quand plusieurs trajectoires réalistes sont considérées simultanément, par exemple au poignet gauche et au poignet droit, afin de vérifier si des trajectoires plus nombreuses (ou plus longues) peuvent offrir une meilleure discrimination. Les résultats de cette expérimentation ont été que pour les trois séquences, les activités sont plus élevées lorsqu'on considère 2 articulations. Plus précisément, elles sont deux fois plus élevées comme prévu. Pourtant, comme l'activité augmente avec le nombre de populations, un seuil variable (t_h) doit être considéré comme fonction du nombre d'articulations afin de parvenir à un meilleur taux de discrimination. Ainsi, si le seuil est t_h pour 1 articulation (ou trajectoire), il devrait être $2t_h$ pour deux articulations afin d'obtenir un taux de discrimination deux fois supérieur.

Le dernier aspect que nous évaluons est la dépendance à l'orientation qui est caractéristique de la discrimination de séquences visuelles. Dans cette expérience, nous montrons comment une

rotation dans le plan de la caméra, donc une rotation 2D de la figure 10, affecte la discrimination d'une population. Nous considérons seulement la population "applaudir", et nous la faisons tourner de $0\,°, 5\,°, 10\,°, 15\,°, 20\,°$ autour de la caméra pour trois entrées possibles, "applaudir", "combattre"et "saluer". Les résultats, où à chaque rotation les trois entrées sont utilisées, mais seule une population est considérée (m_1), montrent une faible tolérance aux rotations. Plus précisément, nous avons remarqué que l'activité moyenne est plus grande pour la population "applaudir" pour les autres entrées à 0° et 5° degrés de rotation, ensuite les stimuli deviennent indiscernables .

Influence de l'extraction de caractéristiques visuelles

Même si nous montrons que notre modèle peut discriminer des trajectoires réalistes (d'articulations) associées au mouvement humain, l'entrée n'est pas la séquence visuelle réelle, mais directement les trajectoires. Une des questions que nous étudions dans cette thèse est la nature exacte des caractéristiques visuelles que pourrait utiliser le cerveau pour effectuer la discrimination de séquences en utilisant des informations de mouvement local dans un cadre distribué, c'est à dire dans le cadre maintenant donné par notre modèle de discrimination, lequel permet ainsi de comparer différents types de caractéristiques visuelles. La caractéristique visuelle la plus évidente consiste à considérer directement le flux optique brut, où l'information spatiale de base est présente, comme certains auteurs l'ont proposé [EMVK09]. Une autre option est que des motifs locaux de mouvement pourraient constituer un élément clé pour effectuer la discrimination [CG05]. Nous expliquons d'abord ce que sont ces motifs locaux de mouvement et comment les calculer, de manière à comparer la discrimination qui peut être réalisée en utilisant soit le flux optique brut soit les motifs locaux de mouvement.

- **Motifs locaux de mouvement** Le calcul des motifs locaux de mouvement trouve de fortes inspirations biologiques dans la mesure où ce type d'opérateurs a été signalé dans l'aire MST, et des lésions de cette aire semblent réduire les taux de reconnaissance de mouvements biologiques [PM94]. L'aire MST est activée lorsque certaines formes particulières peuvent être observées dans le flux optique, comme des translations, des rotations ou des discontinuités. Il a été soutenu [CG05] que ces opérateurs peuvent décrire le flux en termes de ce qui rend une certaine partie du volume spatio-temporel plus importante qu'une autre (en termes d'informations). Sur la base de cette idée, nous précisons le type d'information que nous considérons comme étant le pourcentage de la variance totale entre les petites zones du champ visuel, et nous recherchons les motifs optiques locaux les plus représentatifs.
  Afin d'estimer ces motifs, nous appliquons une analyse en composantes principales (ACP) sur le flux optique (voir plus de détails sur cette technique dans l'annexe C.1) pour chercher les motifs locaux de mouvement les plus pertinents (représentatifs de la plus grande part de la variance). En même temps, nous vérifions si l'analyse est indépendante du type de séquence (mouvement humain ou autre) et de la technique d'extraction du flux optique. Pour calculer l'ACP, nous considérons les vecteurs de données fournis par le flux optique observé sur de petites imagettes carrées de côté L (en pixels). La dimension des vecteurs est $n = 2L^2$, où le facteur 2 provient du vecteur vitesse (vitesse et direction). Nous construisons des imagettes avec 50 % de chevauchement de champs récepteurs. Ce chevauchement ne change pas la dimension mais le nombre de vecteurs disponibles, la dimension dépendant seulement de la taille du champ récepteur (L).
  Les expériences que nous présentons pour étudier les motifs locaux de mouvement considèrent trois bases de données : la base ICCV [MNCG01], la base KTH [SLC04] et notre

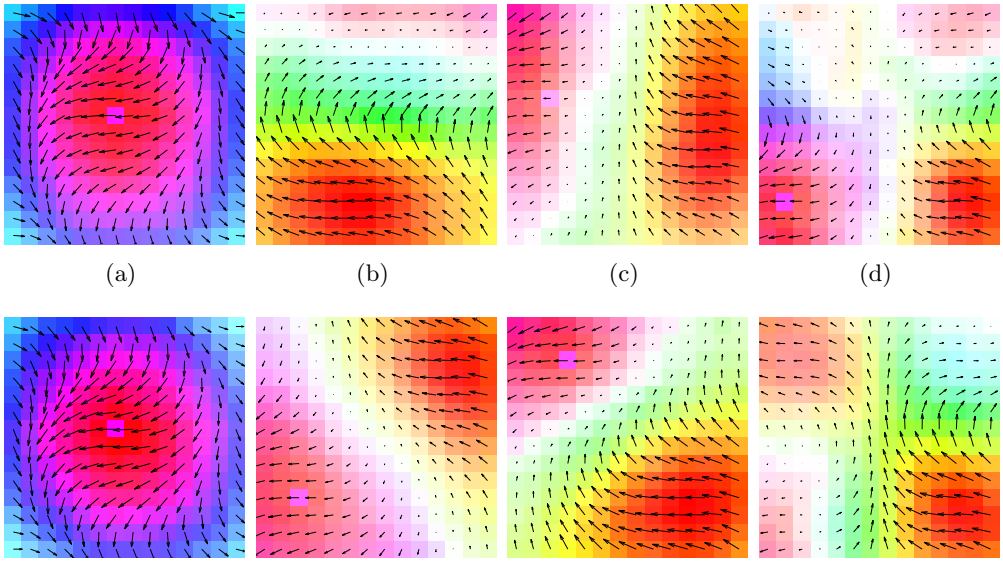
propre base de données de mouvements humains (construite au Loria). La première base de données contient une séquence d'objets rigides en mouvement où le flux optique "parfait" est connu. La base de données KTH consiste en une série d'actions humaines (boxer, applaudir, saluer), avec différents acteurs et caméras où le flux optique véritable n'est pas disponible. Notre base de données consiste en l'animation de trois mouvements humains, les mêmes que la base de données KTH, mais sous la forme d'animations (moins bruitées que les séquences réelles de KTH).

Les résultats de nos expériences montrent que les motifs locaux calculés sur le flux optique sont invariants à la technique d'extraction du flux optique (en comparant l'approche différentielle et le flux optique véritable), comme nous l'expliquons dans l'annexe C.2.1. Donc, même si les techniques d'extraction du flux optique fournissent des résultats bruités, la décomposition obtenue par ACP est similaire. Ce résultat peut être retrouvé, non seulement pour des séquences simples comme dans la base de données ICCV, mais aussi sur des séquences de mouvements humains qu'elles soient réelles (base de données KTH) ou animées (notre base), où les composants ACP ne sont pas très différents, voir figure 12. Un deuxième résultat de cette analyse est que certains composants (ou motifs locaux de mouvement) peuvent être récupérés dans des séquences de mouvements biologiques (KTH ou notre propre base de données), comme dans la figure 12 (a), mais aussi dans des séquences de mouvements non-biologiques (base de données ICCV), voir l'annexe C.1 pour plus de détails. Ce dernier résultat nous permet d'affirmer que si la discrimination de séquences utilise les motifs locaux de mouvement, alors cette analyse n'est pas strictement limitée au mouvement biologique, puisque la base de données ICCV comprend des rotations, des translations et des mouvements centripètes et centrifuges, donc une très grande variété de types de flux optique.

– **Flux optique vs motifs locaux de mouvement**

Dans cette partie de notre travail, nous présentons les résultats de discrimination pour les séquences "applaudir", "combattre" et "saluer", d'abord en utilisant le signal de flux optique brut, puis les motifs locaux de mouvement, en considérant 3 populations différentes pour représenter les 3 mouvements dans notre base. Contrairement à l'expérience précédente qui utilisait les données directes de la trajectoire, dans cette expérience l'entrée est la séquence visuelle elle-même, soit sous la forme du flux optique extrait ou sous la forme de l'information extraite par les motifs locaux de mouvement. Notre résultat principal est de montrer que notre modèle peut utiliser les deux types d'entrées, mais comme le flux optique est trop bruité il donne des résultats de classification faible (la différence entre l'activité totale des différentes populations tend à être petite). Un deuxième aspect est que pour la même population il n'est pas possible de discriminer les différents mouvements (i.e. dire si une entrée donnée est le motif codé ou non), parce que chaque séquence a sa propre quantité de mouvement total, par exemple la séquence "saluer" contient beaucoup plus d'information de mouvement que la séquence "applaudir". Compte tenu de cela, nous explorons l'utilisation de motifs locaux de mouvement, qui offrent une réponse efficace à la réduction du bruit et à la normalisation à effectuer sur le flux, permettant ainsi d'obtenir en moyenne (pour les 3 séquences) une activité totale $1,3$ fois plus importante pour l'entrée correspondante (sur la base de la trajectoire d'une seule articulation).

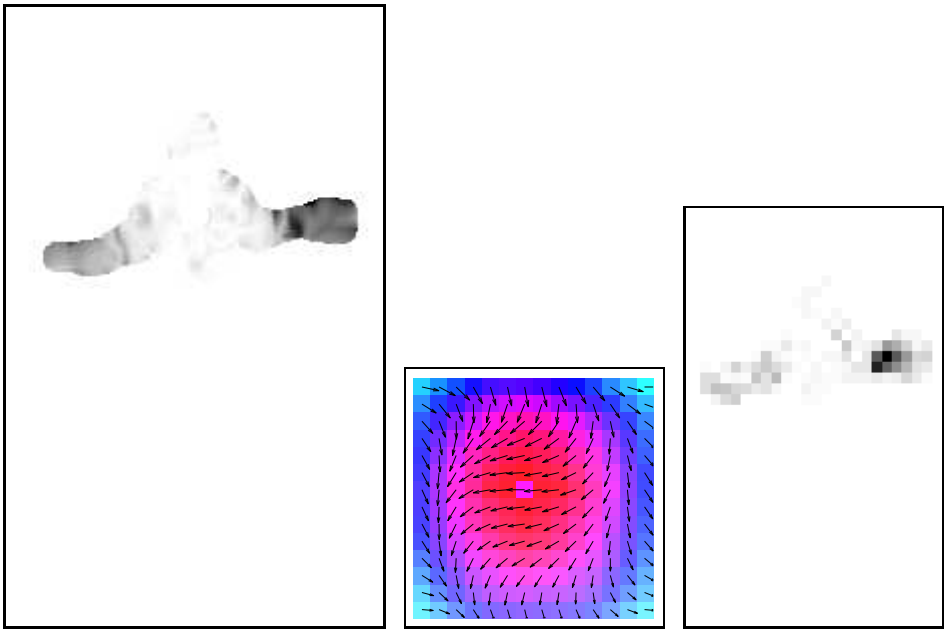
Pour résumer, nous montrons que l'architecture distribuée que nous proposons pour la discrimination des séquences spatio-temporelles est efficace et reste possible à utiliser avec des séquences réelles. Dans le même temps, nous reproduisons un ensemble de propriétés observées de la reconnaissance humaine de mouvements biologiques comme la catégorisation rapide, et l'invariance partielle aux rotations. Nous remarquons aussi que notre modèle, par construction, ne



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

(e) (f) (g) (h)

FIGURE 12 – Les quatre premiers éléments de l'ACP pour la base KTH (première rangée) et la base Loria (deuxième rangée). Dans les deux bases de données, il y a 3 séquences, la taille de l'opérateur est de 14×14 pixels avec 50 % de chevauchement, ce choix étant basé sur les paramètres associés à la zone MST [CG05], en utilisant le flux optique à partir de 18 images. Les actions dans les deux bases de données sont les mêmes, à l'exception de "combattre" où le point de vue est différent.



(a) (b) (c)

FIGURE 13 – (a) Valeur absolue du flux optique, des valeurs plus foncées représentent des valeurs plus élevées, donc les zones de déplacement plus rapide. (b) Composante de l'ACP, 14 ×14 pixels. (c) Projection du flux optique sur la composante de l'ACP décrite en (b), l'image est réduite par un facteur de 7.

peut expliquer les expériences menées sur les stimuli aléatoires PL ou ISI, car une perturbation dans la structure locale (les trajectoires) va interférer avec la discrimination, ce qui n'est pas le cas chez les humains. Nous soutenons l'idée que ce problème peut se poser, parce que notre modèle est au mieux incomplet, et qu'il ne considère aucune information de forme ou information prémotrice pour le rendre plus robuste à ce genre de perturbation. Les résultats que nous présentons relatifs à la caractéristique visuelle précise permettant d'effectuer la classification montrent que des opérateurs locaux de flux optique fournissent non seulement un meilleur taux de discrimination (i.e. rendent le système plus robuste), mais aussi une réduction de la dimensionnalité du problème (le champ récepteur comprend une fenêtre de 14×14 pixels, donc une réduction par un facteur de 7).

Nous concluons, au vu de nos résultats expérimentaux, que les motifs locaux de mouvement utilisés conjointement avec notre modèle ACNFT, fournissent un mécanisme fonctionnel réparti pour la classification des séquences temporelles, et plus particulièrement des séquences de mouvements humains comme nous l'avons montré. Au total, cette architecture distribuée satisfait les propriétés de discrimination rapide, de discrimination incrémentale, et d'invariance partielle par rotation. Ces résultats soulignent l'importance des signaux locaux de mouvement pour décrire les séquences visuelles. L'information de mouvement local, combinée avec un modèle de corps implicite (les trajectoires de mouvement local), fournit une hypothèse sur la façon dont la reconnaissance du mouvement biologique peut être effectuée dans le cerveau : par des populations de descripteurs locaux en conjonction avec les indices globaux (modèle de corps).

III Conclusion

Dans ce travail nous avons étudié la perception du mouvement à partir d'une perspective distribuée, en comparant les techniques de vision par ordinateur et la biologie. Dans la première partie, nous nous concentrons sur deux problèmes de l'estimation locale du mouvement : le problème d'ouverture et celui de l'échantillonnage de la vitesse. Nous montrons qu'il n'est pas suffisant d'effectuer des calculs locaux, une intégration doit être réalisée entre les populations de détecteurs locaux, afin de trouver l'estimation correcte grâce à la dynamique des populations d'unités. Nous comparons nos techniques avec des méthodes en vision par ordinateur et en biologie pour montrer que, dans le cas du problème d'ouverture, donner une sémantique à ce que chaque unité calcule (la direction du mouvement) permet de combiner des informations, dans certains cas de manière plus rapide, mais dans l'ensemble de manière plus robuste lorsque l'estimation initiale est bruitée. Le résultat précédent peut aussi être considéré comme une interprétation fonctionnelle de la structure des champs récepteurs des neurones dans MT, où l'interaction est excitatrice entre orientations similaires, mais inhibitrice pour des orientations opposées de vitesse. Nous explorons également la détection multi-échelle de la vitesse, où nous établissons le lien entre l'erreur relative et la discrimination de la vitesse, des métriques issues respectivement des domaines de la vision par ordinateur et de la biologie. Dans cette perspective, nous montrons que la prise en compte d'une combinaison moyenne pondérée sur la population avec un échantillonnage à échelle logarithmique est suffisante pour obtenir les courbes quasi constantes de discrimination de vitesse observées chez les humains. Par ailleurs, l'erreur relative précise peut être contrôlée par le degré de chevauchement entre les échelles. Lorsqu'on les compare avec des algorithmes sériels de vision par ordinateur pour l'estimation multi-échelle de la vitesse, ce type d'architecture parallèle permet d'éviter une accumulation d'erreur en évaluant simultanément à toutes les échelles.

La deuxième partie de cette thèse est en lien étroit avec la question de la précision requise en vision primaire pour réaliser des tâches de vision cognitive. En effet si l'estimation locale du mouvement nécessite du temps, et si ce temps est plus grand que celui requis par certaines tâches cognitives supérieures, telles que la reconnaissance du mouvement biologique, alors ce n'est pas l'orientation ou l'amplitude précise du mouvement qui constitue l'information essentielle dont le cerveau a besoin pour coder des séquences complexes. Partant de cette idée, nous avons étudié différentes expériences biologiques et psychophysiques, où nous concluons que la structure spatiale du mouvement semble être le facteur déterminant pour coder des séquences visuelles complexes, comme montré par des expériences telles que les stimuli PL. Suivant cette hypothèse, nous construisons un modèle capable de coder n'importe quel ensemble de trajectoires locales, en faisant l'hypothèse que cette décomposition est possible, et nous proposons des critères analytiques simples pour définir les paramètres des populations d'unités, avant d'effectuer des tests sur des séquences simples. Après cela, nous évaluons notre modèle sur des mouvements réels, en montrant que plusieurs des propriétés de la reconnaissance humaine de formes (dans le cas du mouvement biologique) peuvent être retrouvées dans notre modèle, telles que : la discrimination rapide, la tolérance à la déformation temporelle, la discrimination incrémentale, la reconnaissance dépendante à la vue. Enfin, nous évaluons la capacité de notre modèle à coder des séquences visuelles complexes, en utilisant comme entrée soit le flux optique brut, soit des opérateurs intermédiaires de flux optique. Nous constatons que, même si nous pouvons coder les séquences en utilisant le flux optique brut, l'utilisation de motifs locaux de mouvement présente de nombreux avantages, tels que l'augmentation de l'invariance aux rotations, mais surtout l'amélioration de la robustesse au bruit du codage global basé sur notre modèle. Celui-ci, en plus d'une architecture fonctionnelle du point de vue de la vision par ordinateur, présente une hypothèse sur la façon dont les séquences visuelles peuvent être codées dans le cerveau, au moyen de populations de neurones qui codent la structure spatiale du mouvement. Le mouvement local y joue un rôle crucial, mais l'information spatiale pourrait être utile également (comme cela a été observé). Cette idée présente une alternative à l'hypothèse dominante qui a proposé que nous puissions avoir des neurones de "vues" dans l'aire STS/FBA, où chaque instant d'une séquence visuelle serait explicitement codé. Dans le même temps, nous montrons que le rôle fonctionnel des opérateurs locaux de flux optique dans l'aire MST peut être interprété comme une étape de débruitage et de normalisation, afin de renforcer le système en lui-même. Nous vérifions cependant que le flux optique brut peut être néanmoins utilisé comme entrée, mais la capacité de discrimination est alors beaucoup plus sensible au bruit, comme cela a été observé lorsque l'aire MST présente des lésions.

Les principales contributions de cette thèse sont les modèles proposés pour la détection du mouvement et la classification des formes. Nous considérons les tâches de vision primaire dans la première partie, où nous proposons une population distribuée d'unités qui peuvent fournir avec succès une réponse cohérente et une portée de détection plus large que tout détecteur particulier, en utilisant dans les deux cas une population distribuée d'estimateurs locaux de vitesse. Dans la deuxième partie, nous définissons un modèle distribué de reconnaissance de formes capable de coder des séquences de mouvement complexes en utilisant des caractéristiques visuelles locales, où la population connaît la structure locale du mouvement (modèle de corps implicite dans le cas du mouvement biologique). Dans le même temps ce dernier modèle présente une hypothèse sur la façon dont la reconnaissance des formes visuelles dans le cerveau peut être effectuée, non pas en sauvegardant des vues statiques (ou patrons) du mouvement réalisé, mais en enregistrant la dynamique des caractéristiques visuelles locales. Enfin, nous montrons que les motifs locaux de mouvement peuvent aider dans la tâche de discrimination en débruitant le flux optique et en améliorant l'invariance à la rotation et à l'échelle, du moins dans l'architecture proposée pour

coder des séquences de mouvement complexes.

Part I

Early Vision

1 Vision and motion

Contents

In this Chapter, we will introduce basic concepts from computer vision and biology related to the capture and processing of visual signals. We will start by explaining how digital cameras capture and encode images and then continue with an introduction about how animals, specifically primates, deal with visual stimuli. One of the objectives is to explain the main definitions and concepts of perspectives, to point out the differences by the end of the chapter. We will also show how the same fundamental problems eventually arise and how very different solutions exist under both computer vision and biology perspectives. This parallel will be our starting point to study the visual information processing in the brain.

1.1 Computer vision

Before introducing the relevant aspects of computer vision, we present a brief historical context to situate the reader.

Computer Sciences is in general a rather new area of research compared to physics or mathematics. We may find the roots of the subject in mechanical computing machines as early as in the 18^{th} century [Cre93], but it was not until the 20^{th} century that what we nowadays call computer science took form. Specially important where the works of Von Neumann, Turing

and Shannon [Cre93] who among others, in the mid-20^{th} century established the foundations of computer science and information theory used in almost all computers and networks today.

Since the beginning of what we understand as computer science in the 40', an almost immediate challenge was to make computers as powerful as humans' brains, even Turing himself developed a test to check if a given system was as smart as a human (the Turing Test). This area of research took the name of artificial intelligence, and had a period of enthusiasm between the 60'-70', after the works of McCulloch, Pitts and Hebb on neural networks in 1950, Perceptrons in 1958 by Rosenblatt and McCarthy's LISP programming language in the same year [Cre93].

At this time, knowledge representation and symbol processing were the main subjects of research. It was not before the 1960s that image processing started to be developed in the works on the "micro worlds" at MIT IA labs [Cre93], but processing images remained an extremely difficult task for a large period. As more and more powerful computers appeared with time, it became feasible to process images and scientists started to take ideas from mathematics for image processing, geometry, animation. This represented the beginnings of computer vision.

Many image processing problems have been satisfactorily solved in the last two decades, among others: fingerprint and iris identification, barcode reading and OCR (optical character recognition), but always in very controlled scenarios [Pav00]. Problems dealing with real images in normal scenarios remain extremely difficult to solve. To illustrate this, until the year 2000, there was no effective real-time algorithm to detect faces under real conditions (proposed by Viola and Jones in [VJ01]), which is something extremely easy for humans to do. As another example, although 98% recognition rate is possible these systems are not common, one reason being that the recognition rate drops dramatically under real conditions. This explains initiatives such as the CVPR face recognition contest in 2005 [CVP05]. It is in this context that the computer vision community started to look at biology for ideas to address problems such as robustness and operation under realistic conditions. In this Section, we will briefly explain some basic concepts in image and video processing: starting with the creation and manipulation of digital images, giving special attention to the temporal aspects of video processing.

1.1.1 Capturing the light

The starting point in photography is photosensitive materials (*i.e.* photographic paper) able to change with the light. The light reflects onto objects and then goes through a small hole to the photosensitive material, see Figure 1.1. The way a *3D* world is projected into a *2D* plane can be modeled with a pure geometrical model usually known as Pinhole model [BK08], that has been successfully used in the mathematical modeling of image projection. The abstract camera that we describe has at least two parameters: the focal distance (f), and the cutting distance (z_0) depend on the information we want to emphasize in the image (closer or more distant objects for example).

The projective model does not explain light transformation into a digital signal. The light in fact has different wavelengths that define what we know as colors. We cannot see all the colors or wavelengths; for example, infrared and ultraviolet bands are not visible for humans, but many birds have ultraviolet vision, and many shrimps have also infrared vision [CW03]. In digital photography, an array of sensors for each color rather than a photographic material sensitive to light is available in each camera. Common cameras have RGB sensors (Red, Green, Blue), and combinations of these sensors encode for different colors to achieve, in most available hardware, 2^{8+8+8} or around 16 millions possible colors. In color processing there are also several parameters; we can use one (gray scale vision) or several detectors at different wavelengths (color vision), the election of the right configuration evidently depending on the task.

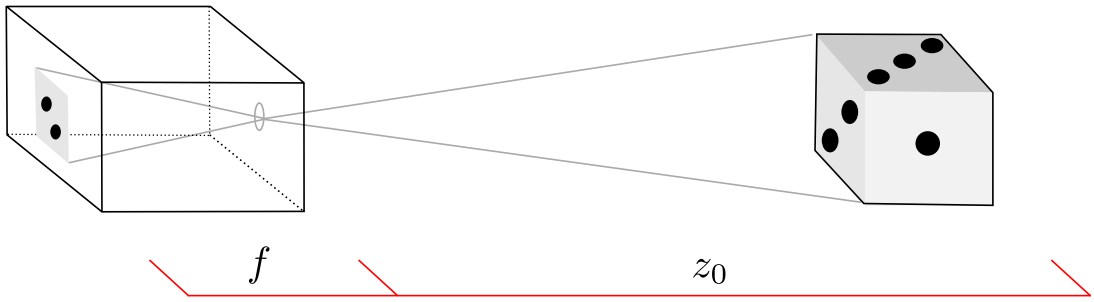


FIGURE 1.1 – Pinhole Model diagram.

After the light is transformed and digitized, the information requires being stored. Most cameras generate WxHx3 bytes of data (3 bytes for RGB color-coding) where images have width W and height H. If the images are compressed (in mpeg, avi or other formats) we consider them as a series of raw images anyhow since we can always derive this series from the compressed format. Nowadays, most current digital video cameras work at $W = 640$ and $H = 480$ or higher resolutions with uniform spatial sampling and time sampling around 30 Hz in non-specialized hardware and 100Hz or more in professional devices. It means that the visual flow of information in a regular camera will be around 90M Bytes of information by second in a very conservative estimation. Real-time operations (< 20ms) require then quite rapid algorithms even with today's computers and, very often, specialized hardware such as GPU or FPGA.

In each one of the 3 aspects we have mentioned, light and how it is digitized, there are parameters. For the Pin-Hole model it is the distance to the object (z_0, f), for the color it is the number of channels (gray scale, two or three colors), for the spatial sampling it is the density of the sensors and their distribution. By consequence, any camera in realistic conditions, implicitly or explicitly has to control these parameters, but if we study image operations that require as few as possible parameters to control, we must simplify the task. Among the features that can be extracted over images, like edges, movement, forms, we focus on one feature that is inherently robust to the camera parameters: motion. Motion has the advantage to be computable even over gray scale images, without depth information and in different spatial sampling, and if more information is available, the movement detection can be improved. In addition, it describes the evolution of images in a sequence, including temporal information. We will now review how we can compute motion from a sequence of images, or the optical flow of a sequence of images.

1.1.2 Feature extraction: the optical flow

Until now we have studied only one image at a time, but including the evolution in time, taken at some temporal sampling rate, delivers a series of images $I(x, y, t)$ or what we commonly know as a "video". In this series of images, we perceive changes, and thinking in terms of the change of position for each pixel, we can define at each time instant the vector of change $\vec{u}(x, y, t)$, also known as the optical flow. Nevertheless, why should we focus on motion detection if static images contain large amounts of information? Indeed, this is precisely why motion is an interesting feature to compute: as only a few elements in the visual scenes are most often moving they are usually more meaningful, thus providing a powerful way to reduce the visual information to process.

As we noticed, the visual flow of information could deliver abundant information about a scene and we argue that in general temporal evolution helps to extract information, but we have

not explained how to actually compute the optical flow, or $\vec{u}(x, y, t)$ from $I(x, y, t)$. We will now explain several strategies to actually compute $\vec{u}(x, y, t)$.

Differential Approach

Many optical flow extraction methods are based on the assumption of brightness conservation [1], that is,

$$\frac{dI(x, y, t)}{dt} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{1.1}$$

where $\vec{u} = (u, v)$ is the velocity vector. A well known technique following this approach is the Lucas & Kanade algorithm [Luc85], that minimizes the following cost function in a small fixed region Ω,

$$\vec{u} = \arg\min_{\vec{u}} \left\{ \sum_{\vec{x}\in\Omega} W^2(\vec{x}) \left[\nabla I(\vec{x}, t)\cdot \vec{v} + \frac{\partial I}{\partial t}(\vec{x}, t) \right]^2 \right\} \tag{1.2}$$

where W is a two-dimensional Gaussian function used to give more influence to the central points and Ω is a square region of a few pixels. This minimization estimates $\vec{v}$ with sub-pixel precision after a few iterations. This method achieves a good optical flow estimation in regions where $|\nabla I(\vec{x}, t)| > 0$, such as corners [Sim99].

Correlation approach

In the most basic versions of the correlation-based approach [BFBB94], two consecutive images frames are taken into account. For each pixel, the translation $(\vec{d})$ that maximizes the correlation between a region around the pixel and potential regions in a search area is taken as measure of the displacement at each location. This technique tends to preserve optical flow discontinuities but problems arise with symmetrically textured moving areas, not detecting or wrongly detecting motion directions/intensities. The method can be summarized as a correlation stage followed by a selection of movement:

$$M(\vec{x}; \vec{d}) = W(\vec{x}) * Corr(I(\vec{x}, t), I(\vec{x} + \vec{d}, t + 1)) \tag{1.3}$$

where I stands for the intensity image, W is a weight function, $Corr$ is the correlation function and $*$ is a convolution. A simple instance of this family is the SAD (Sum of Absolute Difference) algorithm [BFBB94], in which $W(\vec{x}) = W = 1$ and $Corr(A, B) = |A - B|$:

$$M(\vec{x}; \vec{d}) = W * |I(\vec{x}) - I(\vec{x} + \vec{d}, t + 1)| \tag{1.4}$$

M is the target to be maximized at each pixel location as a function of $\vec{d}$ for the SAD algorithm. More elaborated versions include the previous use of directional filters (to select borders), to use different scales to achieve a wider range of displacement, of fitting curves with the correlation outputs for sub-pixel precision, or of simultaneous correlation of several pairs of frames to smooth out results [BFBB94].

1. The same point viewed from a camera at two time instant (or several cameras) does not change its brightness (Lambertian Surfaces)

Frequency approach

Extraction of the optical flow in the frequency domain [Hee87] is based on the outputs of filters tuned, in the Fourier domain, to different speeds. Two basic ideas in this approach are:

- The power spectrum for any image is assumed as flat.
- A rigid translation in the space-time domain gives a rotation of the original power spectra plane (this is another way to assume implicitly the brightness constraint).

$$\mathcal{F}\{I(x-ut, y-vt, t)\} = c(w_x, w_y, w_t + uw_x + vw_y) \tag{1.5}$$

Equation 1.5 stands for the second idea, where $\mathcal{F}$ is the Fourier transform and c is the Fourier transform of a static image sequence. This expression gives a prediction of how every possible response in the Fourier domain will be for different velocities, and hence a way to compare an observed versus estimated frequency response. This method smooths out results, and consequently optical flow discontinuities are diminished. Interestingly, recordings of motion-tuned neurons on different animals have shown the frequency-tuned nature of brain neurons making this model a biologically plausible method for perception of motion. Also interesting is that given a desired precision in the speed domain, a fixed set of filters can be configured, so that any response can be directly interpolated [AS08]. It is also important to mention the algorithm proposed by Heeger [Hee87], that we can briefly summarize as,

$$K_{\vec{w}_i} = G_{\vec{\sigma}}(x, y, t) * \left(L^2_{\vec{w}_i,0} + L^2_{\vec{w}_i,\pi/2}\right) \tag{1.6}$$

$$L_{\vec{w}_i,\phi}(x, y, t) = Ga_{\vec{w}_i,\phi}(x, y, t) * DoG_{\sigma_1,\sigma_2}(x, y, t) * I(x, y, t) \tag{1.7}$$

$$Ga_{\vec{w}_i,\phi}(x, y, t) = G_{\vec{\sigma}}(x, y, t)\cos(2\pi\{w_x x + w_y y\}) \tag{1.8}$$

where the $*$ operator stands for the convolution. The oriented filter (L) in Eq. 1.7 is a Gabor (Ga) with a Difference of Gaussian operation (DoG), where the Gabor filtering has a center frequency $\vec{w}_i = (w_x, w_y)$ and a spatial support, given by $\vec{\sigma}$. Equation 1.6 is a quadrature pair of the filter L smoothed by a low-pass Gaussian filter G. Other authors have proposed derivatives of Gaussian functions instead of Gabor Filters to span the frequency domain [SH98].

Heeger [Hee87] uses as initial processing 3D Gabor energy filters, set according to 12 different spatiotemporal orientations that define a profile of activation for each given location (12 version of Eq. 1.7). Then he analyzes the filter outputs combined in a quadrature pair (Eq. 1.7) as a code for the searched velocity. This code is precomputed for a set of predefined velocities to compare the predefined and the obtained response of Eq. 1.7 by the least squares technique and retrieve the optical flow.

1.1.3 Constraints and implementation

Why so many methods exist to compute the optical flow? Simply because it is an ill-posed problem, the solution at each pixel (x, y) may not be unique, at least in the way we have presented the problem. If we rewrite Eq. 1.1 as the product between the velocity and the gradient of the image,

$$\nabla I(x, y, t) \cdot \vec{v} = -\frac{\partial I}{\partial t} \tag{1.9}$$

we can see that only the component perpendicular to the edge can be estimated. This can be also found numerically if we express Eq. 1.2 in a closed form [Luc85],

$$\vec{v} = \begin{bmatrix} W^2 * I_x^2 & W^2 * (I_x I_y) \\ W^2 * (I_x I_y) & W^2 * I_y^2 \end{bmatrix}^{-1} \begin{pmatrix} W^2 * (I_x I_t) \\ W^2 * (I_y I_t) \end{pmatrix} \tag{1.10}$$

where I_x, I_y and I_t are the partial derivatives, and $*$ the convolution operator. As this algorithm requires inverting a matrix, if one of the spatial derivatives is zero or even close to zero, the solution will not be unique, thus it correctly works close to corners. This may be a problem only with the differential approach, but in fact correlation and frequency based approaches compute a likeliness function for each possible speed (selection of speed), so that in these algorithms the problem can be seen as a flat likeliness function [BFBB94]. A fundamental problem in computer vision lies behind this: if you look at local motion through a small hole, you are only able to perceive the movement that is perpendicular to the edge you are looking at, see Figure 1.2. Therefore, there is a unique solution in corners, but an infinite set of solutions along edges.

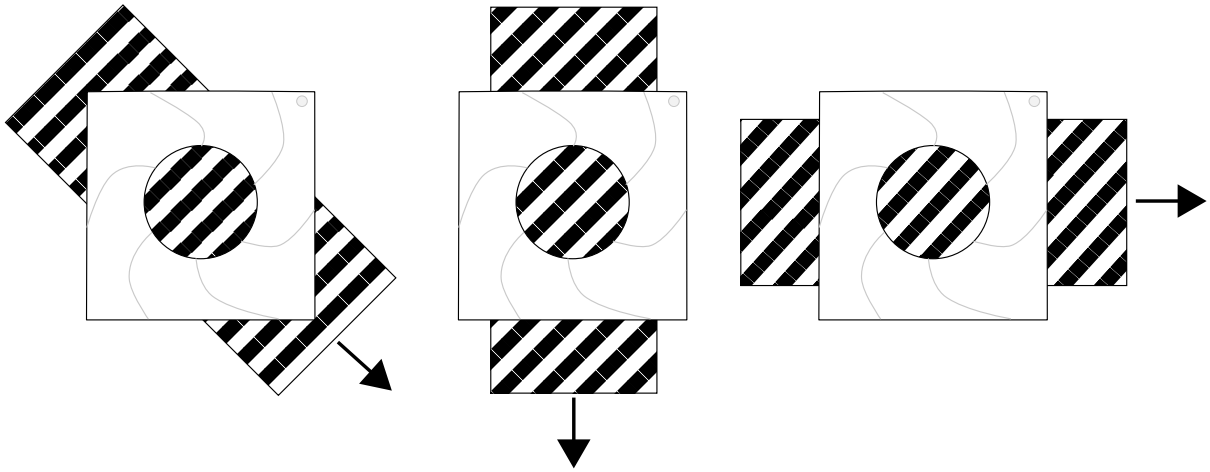


FIGURE 1.2 – The aperture problem. Looking from the aperture, all three movements are identical (reproduced and modified from [AB85]).

Another problem is that all the methods we have presented have in common the fixed range of speed they are able to detect, about the size of the convolution window, for example the size of W in the differential and correlation approaches and the size of the spatial support $\vec{\sigma}$ of the band-pass in the frequency approach. In general the range of detected speeds is fixed and uniform, from v_1 to v_2 and with an absolute error within this range. A common solution to this problem is to work sequentially in several scales of movement detection or to perform a multi-scale analysis with the drawback of slowing down the process as more scales are considered, and with an implicit accumulation of error.

1.2 Biology

Day and night, in people crowded conditions or natural environments, with ambiguous and partial visual information coming from every possible direction, we are still able to identify danger signals or to identify humans. What is so special about our eyes and brain that makes this possible? Biologists have been searching for decades [Ram91, CW03], but we are still very far from understanding the mechanism of the brain and how it works. We will start this section by

reviewing classic and well-established anatomic and physiological knowledge about the brain processing of visual information, and more recent studies. Our description of the visual processing in the brain will show how things are exquisitely complicated and that the straightforward comparison with computer vision will be difficult to establish. Accordingly, the parallel between visual processing and computer vision will be presented in 1.3, to the extent where it remains possible. By the end of the chapter we will focus on two specific problems stressing this comparison: the aperture problem and the multi-scale speed detection.

1.2.1 Capturing the light: The Retina

All the visual processing that our brain performs starts in the eyes. The eyes are a wonderful collection of photoreceptors that roughly translate light into electrical current. In humans, there are two kinds of photoreceptors located at the bottom of the retina: cones and rods, see Figure 1.4. The cones are specialized in daylight conditions, having three families for the different wavelengths (yellow 564-580 nm, green 534-545 nm and violet 420-440 nm). On the other hand, rods are specialized in low-light conditions, and they are able to capture even single photons [CW03]. Other species, like birds [VF93], have different (more numerous in birds) families of cones. The spatial distribution of cones and rods is non-homogeneous, with a high-density zone in the center almost without any rod [WS00], see Figure 1.3. Both receptors also present different responses to light changes: cones are rapid but not very sensitive, rods are slow but they can operate in very low light conditions.

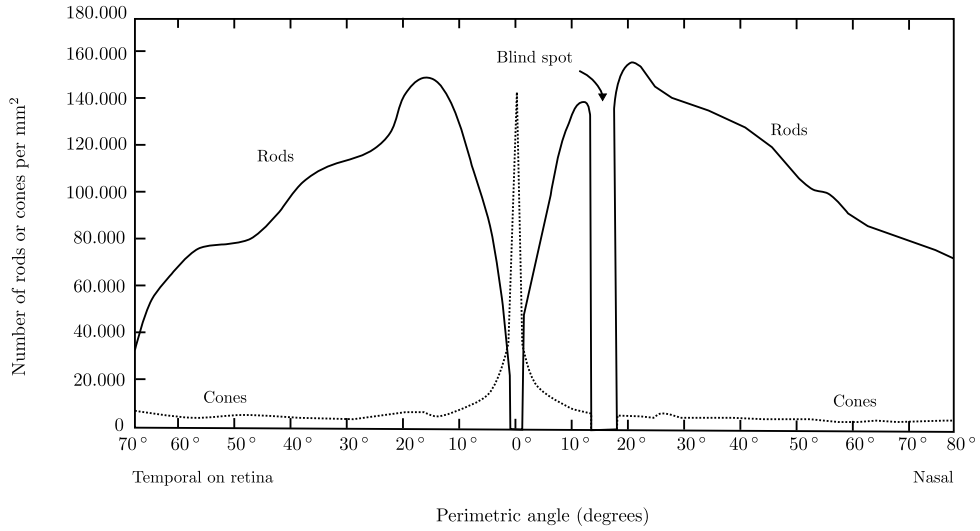


FIGURE 1.3 – Cones and Rods distribution as function of the distance to the optical center (reproduced from [BL98]).

After cones and rods have transformed the light into electrical current, this signal is processed by a series of cell layers in the retina: bipolar, horizontal, amacrin and ganglion cells, see Figure 1.4. This last layer is the output of the retina sending spikes of activity to the primary visual cortex going through the thalamus, specifically the Lateral Geniculate Nucleus (LGN), see Figure 1.5(a). Each layer has numerous sub-types of units [FC07]. Around 10 years ago, biologists were thinking that all ganglion cells were divided only in two categories: midget ganglion cells (projecting into parvocellular pathway) carrying color information in opposition channels

and parasol ganglion cells (projecting into the magnocellular pathway) carrying mostly luminescence information. For 80% of ganglion cells this is true, but we now know [PGS+07] that there are populations specialized in other functions such as detecting object approaching, measuring the light for our internal clock (circadian cycle) and some others which function has not been identified [GM09].

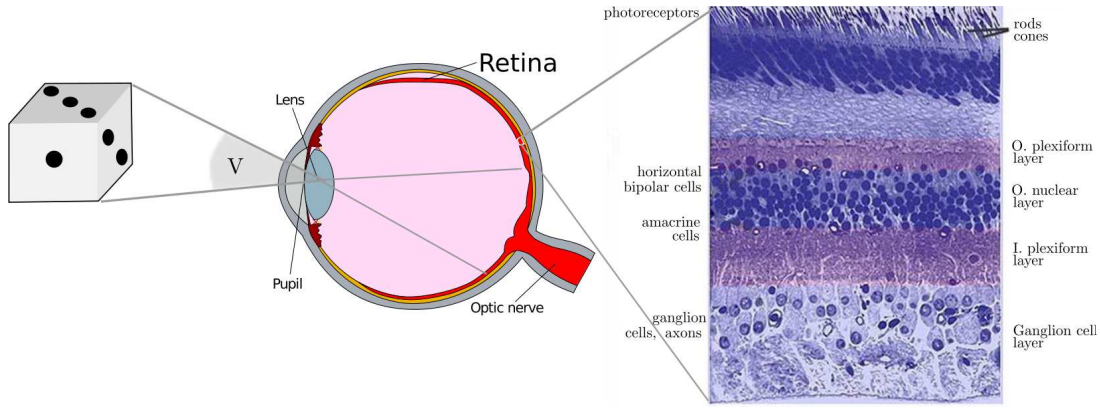


FIGURE 1.4 – The eye (left) and the retinal circuit (right). The retinal image corresponds to a light micrograph of a vertical section through central human retina (reproduced from [HK]).

The high definition zone where most cones are located is called the fovea until around 12° of the visual field, see Figure 1.4. The zone outside the fovea is called periphery, and here the rods present a higher density. Rods and cones are connected to retinal ganglion cells through bipolar cells. Amacrin and horizontal cells contribute to reduce the redundant information, thus enhancing the contrast of the signal (as a first approximation) and combining the signals from cones and rods. Colors are not transmitted directly but in opposition channels: yellow-blue and red-green, the two parts of the P pathway.

1.2.2 Dealing with the visual signal: The primary visual cortex

After the light has been transformed, enhanced and multiplexed into different channels (mainly M and P pathways) it goes through the optical nerve into the LGN [2], an area that relays the information into the primary visual cortex located in the back part of the brain (occipital area, see Figure 1.5).

From the LGN, the visual signal goes to several areas such as V1, V2, and V3 and through different relays into higher areas in the dorsal sense, see green arrow in Figure 1.5. The first researchers to identify the functions of neurons in this area were Hubel and Wiesel [HW62, HW68], that determined that neurons in the primary visual cortex have a small receptive field. Many neurons in the primary visual cortex are sensitive to a small part of the visual field, and, in the case of V1, they are sensitive to light bars in an area smaller than 1 deg^2 [MN87], as it has been measured for the fovea, see Figure 1.6. One model [FM00] for this response is that each V1 neuron takes different spatial arrangements of outputs from the retina, forming a bar at different orientations, phases and speeds (see Figure 1.7). Following the theory of a feed-forward flow of information, neurons in area V5 (or MT) [3] have been associated with larger receptive fields

2. A high percentage of connections to LGN are actually coming from the cortex, but their function is not well understand, but seems to be mainly a modulation signal [CW03]

3. V5 in humans suppose to be equivalent to MT region in macaque monkeys

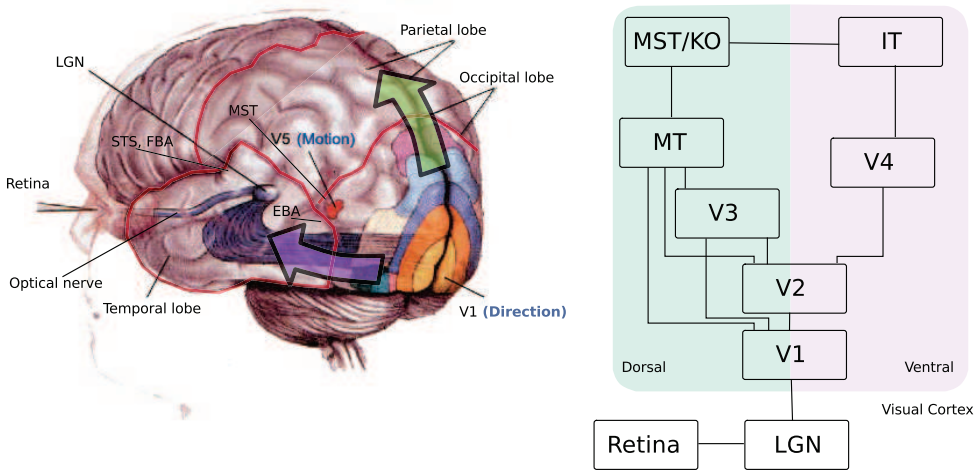


FIGURE 1.5 – (Left) Brain general view illustrating lobes position, some visual areas and the dorsal/ventral pathways (green/purple arrows). (Right) The visual cortex schematic view.

than V1 (4 to 7 times larger [SH98]). In V5, neurons are supposed to have orientation/disparity properties similar to V1 neurons. Current theories claim that area V5 acts as an integrator in space and a segregator (many units in MT are sensitive to disparity) to overcome problems such as the aperture, where it is important to segment objects. This theory is a subject of discussion as a large percentage of units in V1 are end-stopped, and therefore they carry the solution to the aperture problem. A complete discussion about MT function can be found in [BB05].

The LGN-V1-MT circuitry is clearly only one of the many different networks. To give an idea of the extension and richness of these networks, Figure 1.5 (right) presents a diagram of the different brain areas involved in the processing of visual information from the classic literature [CW03]. In this complexity, an idea of two flows of visual information divides the information into two pathways: the ventral pathway for static features and the dorsal pathway for the processing of motion and location in space. Areas V1, V2, V4, IT and F5 dedicated to the processing of static features compose the ventral pathway. Areas V1, MT, MST, KO and STS compose the dorsal or where/how pathway for the processing of motion and location in space. Despite this separation, it appears more and more clear that both pathways are deeply connected and interdependent, for example in the detection of movement based on color [TDA01]. We will continue the study of higher areas in these two paths in Chapter 4, but now we will focus on the local detection of motion and the link with the computation of optical flow in computer vision.

1.2.3 Local detection and speed resolution

As we have seen, computer vision and the brain approaches are radically different in the way the light is interpreted and more specifically movement is detected. To compare both perspectives is difficult as the computation substrate is totally different. In order to compare them clearly, we select only two fundamental problems that are independent of the substrate: the aperture problem and the speed resolution. These problems arise when speed detection is performed by means of local detectors with finite speed detection range which is the case in the brain, as we will show.

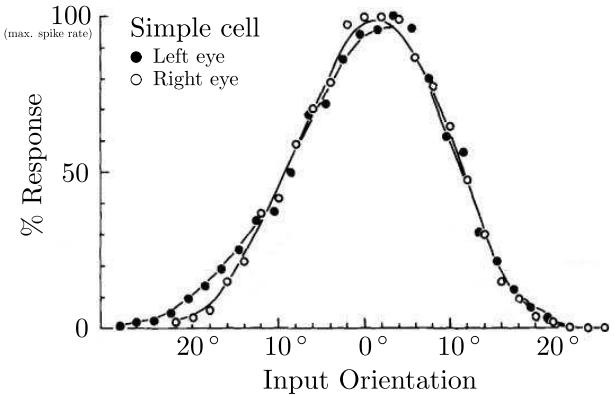


FIGURE 1.6 – Tuning curve in a V1 simple cell from the cat visual cortex (reproduced from [Bis84]). The cell is selective at a given orientation (largest % of activity) and it reduces its activity as a function of the distance from the optimal stimuli.

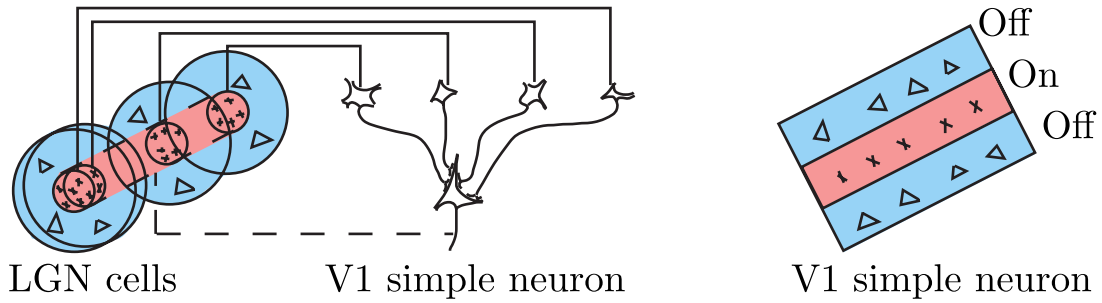


FIGURE 1.7 – The Hubel & Wiesel (1961) model of orientation selectivity in the V1 simple neurons (reproduced from [FM00]).

We have explained how a spatially selected array of retinal ganglion cells output could deliver enough information to build a simple bar detector and this is supposed to be the basic mechanism behind the operation of many neurons in area V1. Speed selectivity implies orientation selectivity but it also requires a certain temporal delay, (as $v = dx/dt$) and this delay is supposed to be given by V1-V5 units, whose combined circuitry provides an oriented speed detector. However, as detection is performed by means of local detectors the aperture problem also arises in the brain operation as we have discussed in section 1.1.3. There are several theories about how the aperture ambiguity is solved in the brain; one of them says that end-point cells (neurons capable to respond to moving corners) are able to communicate their error-free information to other movement detection cells [CCPB03, BB05]. Another theory claims that lateral connections are sufficient to solve the aperture problem (no need for special units) [SH98, BB05] and another theory says that there is an intra-cortical loop between V1-MT areas, which is able to solve the aperture problem by successive larger receptive fields, in a recursive loop [BN04].

Spatial resolution has been extensively studied in the retina, where precise spatial distributions for each kind of sensors are well known, but as we study the visual cortex in the dorsal way (see green arrow in Figure 1.5) the speed detectors distribution is less precise [MVE85]. In higher areas like V5/MST experimental psychology results are more often available, providing a general description of these areas. Studying some of these experimental results in humans and

monkeys [MN84, OCDBM85, NHD05], the range of speed resolution decreases (increasing the lower bound) with the distance to the center of the visual field (fovea), but keeping the relative error constant over the whole visual field (around 5%). This detection range can partially be explained by the retinal sampling distribution, as well as by cortical magnification factor (foveal areas of the retina have a larger area of processing in the visual cortex). Yet the cortical magnification factor does not explain how the speed detector are distributed and combined to allow discrimination of high speeds in the fovea and periphery of the visual field.

1.3 Computer vision and biology

As we have explained, the idea of local computations of images to extract the optical flow helps to provide efficient implementations either in software or specialized hardware for the computation of optical flow, but we have also seen how this type of solutions inevitably comes up against the aperture problem. In computer vision we digitize images and videos into bytes of information, but the brain generates a continuous flow of information, where the representation of information is not completely understood. However, we know that this stream of information is temporally encoded into spikes of activity, where the first spikes seem to contribute to the most relevant information [RVR01]. Despite the differences in the coding schemes for visual information, neurons have been found in the brain with a response profile that can be associated with a local speed detector (areas V1-MT). This similarity motivates the study of the perception of motion in primates; to understand how the brain deals with the fundamental aperture problem. The aperture problem is not by any criteria the only issue when movement is detected by means of local detectors, but it is a good starting point to compare the two approaches.

A second difficulty with the local detection of motion is that it implies a limited range of speed detection. This problem arises in both computer vision and the mammalian brain: how a range of speed greater than the range of one single detector can be constructed. Computer vision commonly proposes a scheme where the same signal has to be analyzed in multiple scales, starting with the highest one to estimate a series of approximations, *e.g.* for the speed 3.14 pixels/frame estimating first 3.0 pixels/frame (rough movement) then in a successive scale 3.14 pixels/frame (fine movement). This idea tries to provide a constant level of absolute error in the speed estimation. The brain seems to have chosen a very different strategy processing multiple scales (temporal and spatial) in parallel as experiences have shown, and somehow combining these estimations [MVB94, NHD05] (higher speeds do not require more time to be estimated), to obtain a wide range of speed detection with constant relative (and not absolute) speed estimation error. To understand the exact difference between the two strategies and the precise neural mechanism to obtain a wide range of estimation requires the study of speed range determination. This problem provides the second element of comparison between computer vision and the brain.

The aperture and multi-scale analysis in the detection of motion are two fundamental problems in vision (although not the only ones) observed in artificial systems and in the brain despite the extremely different computation substrate. Radically different solutions are proposed under each approach, and in the next two chapters of this thesis we will present two models of how the brain could solve these problems in the context of distributed computation. More precisely, we will show how a distributed system can solve these problems (which are not solved by purely local information) and how the dynamics of the population, together with the local information, can provide the solution to the aperture and the multi-scale speed detection problems.

2
The Aperture problem

Contents

The aperture problem is naturally associated with any local detection system for the visual perception of motion because the response of local detectors is ambiguous. Biological systems, such as the primate brain, are able to disambiguate motion detection, although this operation is also performed locally. Several models in computer vision and biology have been proposed to perform this disambiguation, combining the information of close local motion detectors or identifying some special locations where the detection is not ambiguous to propagate this information. Most of these models have the disadvantage of disambiguating in a fixed neighborhood, thus only partially solving the aperture problem.

In this chapter we present a neural model of motion disambiguation, where close detectors share information and, as a result, the activity converges to the true motion response. Here, it is the system dynamics that delivers the answer and there are no special units to identify unambiguous signals. The proposed model consists of two layers of processing: the first one where local detection is the input signal; and the second one where information is combined in a larger spatial neighborhood and a coherent signal is sent back to the first layer through multiplicative feedback. Our model differs from previous works where this feedback architecture have been already proposed, mainly by incorporating lateral inhibition in the combination step to eliminate spurious signals, and in some cases by delivering a more rapid response.

With this model we show that a population of local movement detectors can overcome the aperture problem by means of population dynamics. We also show, in opposition to known

techniques in computer vision, that the range of speed integration is not limited by the size of the receptive field but by the time the system iterates. We validate our model with synthetic and real image sequences and end the chapter with the discussion about the plausibility of our model.

2.1 Aperture problem and disambiguation

In the context of optical flow extraction, many approaches have already been developed to solve the aperture problem. In this section, we are interested in the main models based on local motion detection that are able to deal with the aperture problem, mainly because these models better fit the current knowledge about how biological systems perform motion detection and integration. We start by a short presentation of the aperture problem and discuss several experiments in biology that have inspired related models.

2.1.1 Aperture Problem

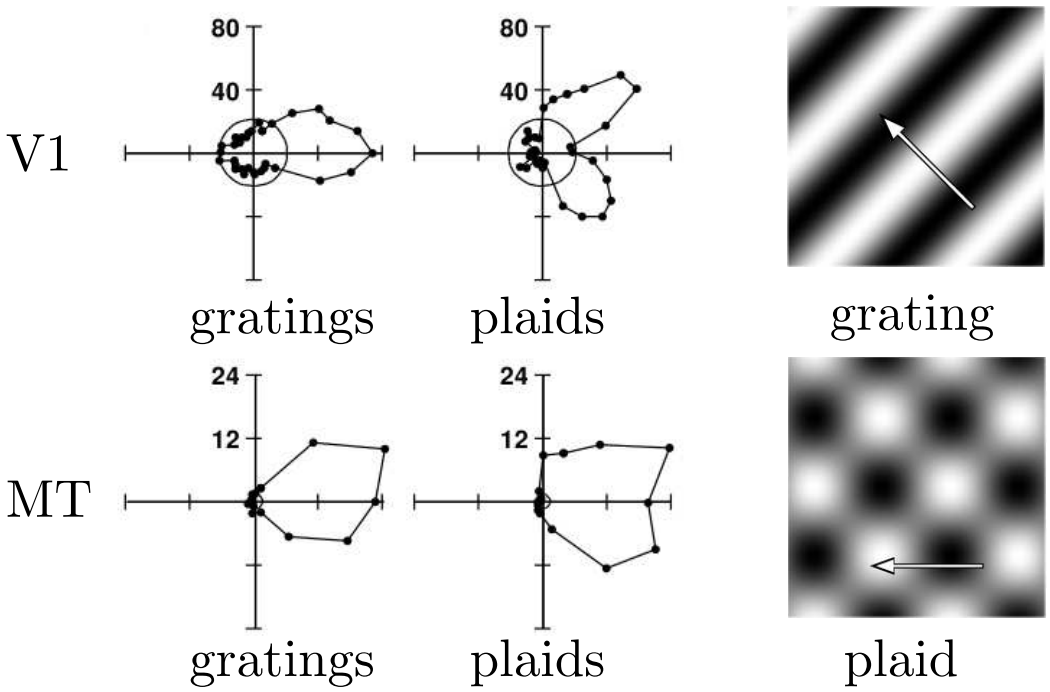
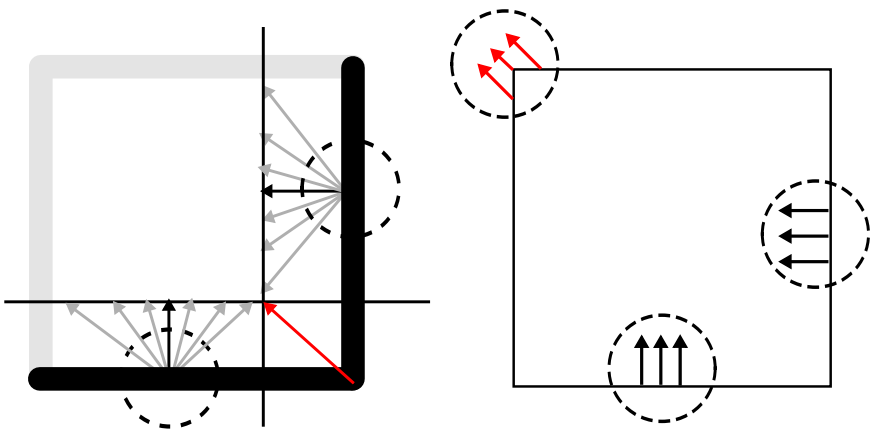


FIGURE 2.1 – Receptive field of V1 and MT neurons for the grating and plaid stimuli. The input stimuli correspond to the angular coordinate and the activity (firing rate) is proportional to the radius. In V1, for example, the biggest activity for a grating can be observed for an input at 0°, where the activity has a firing rate close to 80. The inner circle for each plot represents the spontaneous activity level of each cell. A plaid is made of two gratings, and V1 neurons are selective to one of the components of the grating. Contrastingly, MT neurons are selective to the true motion signal of gratings and plaids. The two cases, grating and plaid, move following the white arrow and correspond to 135° and 180° respectively (from Movshon et al 1986).

The aperture problem always appears when the optical flow is estimated by means of local detectors as we explained earlier in Chapter 1. When we look through a small aperture, it is only the motion component that is perpendicular to the local edge that can be detected. In the visual cortex of primates, V1 neurons have a small receptive field, most of them tuned to a certain spatiotemporal frequency. When the movement is ambiguous within the receptive field of V1 neurons, due to the aperture, neurons activity is sensitive only to a certain component of the movement, see Figure 2.1. Neurons projecting from V1 to MT do not seem to solve the aperture problem [MN96], but units in area MT after a certain time delay (50-75 ms by single units recordings [LPB01]) are activated by the true motion, solving the aperture problem, see Figure 2.1. Similar time latencies have been reported in psychophysical experiments where smooth pursuit of an ambiguous moving target (due to the aperture) presents an initial error that drops after 100 ms or more (MT is directly related to oculomotor control areas associated to smooth pursuit tasks [BFM10]). The receptive field in area MT is also bigger than the receptive field of V1 neurons by a factor of 5 and within this receptive field a complex structure has been reported. It is in this structure that coherent orientations seem to reinforce a given orientation and opposite orientations seem to inhibit a given orientation [LPB01]. The exact shape of this interaction, however, is not well understood.

With respect to possible mechanisms that could explain the previous results, several models have been proposed to explain how the brains solves the aperture problem. It can be noticed that even though a local detector is inherently ambiguous, it is activated by local movements within a limited range, see Figure 2.2(a), so that several detectors may be combined to give unambiguous responses. This combination is called IOC (intersection of constraints, [MAGN85]), as represented by the red arrow in Figure 2.2(a). Another important observation is that local detectors are able to perform directly unambiguous detection for special features such as corners, see Figure 2.2(b).



(a) (b)

FIGURE 2.2 – The aperture problem for a square and its solutions. In (a) the detected movement using two detectors (shown darker) must be combined to obtain the true speed (in red) using the IOC. Whereas in (b) the true speed is directly computed at the corners.

2.1.2 Disambiguation mechanisms

The aperture problem is not particular to a given optical flow algorithm such as the Lukas & Kanade differential approach, Heeger's frequency based approach, correlation strategies such

as SAD or to the movement detection performed in the primate brain. Restraining ourselves to the models based on local detectors, in order to remain biologically plausible, we can distinguish 1D and 2D models to explain how the disambiguation mechanism may work. The distinction is about how information is interpreted because motion detection is always performed in 2D since the optical flow is defined in 2D.

1D detection

Isotropic models have uniform movement detectors, processing each part of the visual field in the same way. The solution to the aperture problem in these models arises from the combination of these local detectors, but there is no special information about corners or junctions. In that sense the detection is only performed perpendicularly to the edge orientation or in 1D. Among these models we can find the "Intersection of constraints" and "vector averaging" algorithms.

Intersection of constraints (IOC). Introduced by Horn et al. [HS81], the IOC idea supposes that a set of close local detectors constrains the space of solutions in such a way that only the true velocity remains. The IOC algorithm, see Figure 2.2(a), was originally proposed in computer vision and later studies in the cat brain by Movshon et al [MAGN85] suggested a neural implementation by combining a set of local detectors as AND detectors, in order to solve the aperture problem. The approach, however correct, is purely theoretical and a neural implementation to support it has not been found. A more plausible implementation was proposed by Simoncelli & Heeger [SH98], where detection is a linear combination of a fixed set of filters to cover the plane of movement in the frequency space. Specifically, the optical flow estimation algorithm proposed by them consisted in 12 different filters with center frequency $\vec{w}_k$, 8 for movement and 4 for static configuration, see Chapter 1 for details. However, when solid moving objects are observed, not plaids or gratings, the solution is propagated to a fix neighborhood, larger than the size of one local detector but that could still be smaller than the size of a given object, thus not completely solving the aperture problem.

Vector average. This method was already mentioned in [MAGN85] and it is a particular simplification to the IOC idea. Vector average uses the idea that when moving gratings form an angle of $\pi/2$ (one grating can be transformed into the other by a $\pi/2$ rotation), the IOC model reduces to the average of the movement of the gratings, a solution that still holds for moving objects. Vector average has an error smaller than $\pi/2$ for right angles objects (like squares) and it provides the right solution in this case. However, the solution is incorrect for objects with corners that are not right angles (like a moving triangle). In this case the error is only smaller than π. The implementation is considerably simpler than the IOC, but as the neighborhood is still fixed the true velocity will not be detected everywhere for solid objects.

2D detection

In the anisotropic (2D) detection, special units are active, or modulations in the interactions among them play a role. In general the goal is to build a "confidence" measure of the local movement detection and to propagate the information from zones of high confidence to zones of lower confidence.

Corner/Junction detection. The report of neurons in areas V1 and MT, that are selective not only to a moving bar but to moving "terminators" such as corners, has initiated an extensive debate about the functional role of such neurons [CCPB03]. Starting from these ideas, other authors [BN07a, CH10] have proposed practical implementations of such a mechanism, where junction detectors are explicitly modeled. The main problem with these detectors is that several

kinds of detectors must be included and combined in order to have an effective detection. For example X and T-junctions to identify true and false terminators (false terminators may arise from overlapping several moving objects).

In the context of computer vision, especially video tracking, corner detection combined with motion detection is often found: instead of computing the optical flow everywhere, the computation is only performed close to corners in points usually called "good features". A simple instance of this approach is to compute $\nabla I(x, y, t)$, and select a few local maxima where to compute the optical flow.

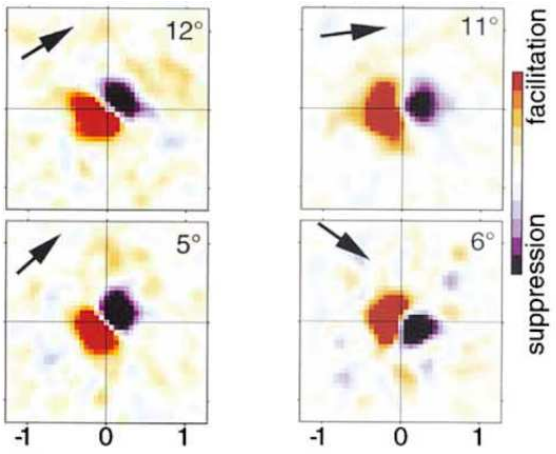
Probabilistic models. In a more artificial approach, Bayesian models have been applied to the disambiguation of motion. These models often use corner/junction detection but within a probabilistic framework. Weiss et al [WA95] express the motion detection in probabilistic terms, not quantifying it as a precise value but as a probability distribution. It is interesting to notice how a confidence function can be actually built: as the aperture problem can be detected as the flatness of a likeliness function, the flatness can be seen as a confidence of measure. For example, the operator $\nabla I(x, y, t)$ can be used as a confidence measure, because it can be associated directly to corners (higher values on these zones).

As we have seen, both approaches (including explicit feature detectors or not) induce algorithms to compute the optical flow, partially solving the aperture problem in most cases. The derived solution is static and as local detectors have a fixed receptive field, these solutions to the aperture problem will not be effective in cases such as a moving square where objects are larger than the receptive field, see Figure 2.2. An interesting idea is to study a system that propagates information in space over time as the one proposed by [BN04] where a recursive mechanism is implemented that can potentially overcome the limitations of a fixed receptive field. The mechanism proposed by [BN04] can be associated to a 1D solution, because there are special units for corners providing a neural implementation of the IOC. This idea is not limited to the aperture problem, but more generally to the problem of propagating a particular solution. It has important support from neurophysiology as feed-back connections have been reported between areas V1 and MT [HPL$^+$98, HJG$^+$01]. Moreover, other studies [OSK03, BFM10, SMM10] have reported that the estimation of speed requires some propagation time, supporting the idea of a diffusive process that require some time to converge, and by consequence supporting the idea of a propagation mechanism.

Other experiences in area MT of primates illustrate another important insight of the motion disambiguation mechanism. Snowden et al [STEA91, BB05] show that the response of MT cells for moving dots is lower when the stimuli are combined with opposite motion (transparent motion). In a more recent work, Livingstone et al show that an architecture where related orientations are reinforced and opposite motion orientation are inhibited may exist within the MT receptive field [LPB01], see Figure 2.3. This behavior was reported in wide range of eccentricities and within the receptive field of MT neurons. The described receptive fields have been reported also to change over time, but we notice that in all cases there is facilitation and inhibition as a function of the preferred orientation of the MT neuron, despite the eccentricity or the precise timing.

2.2 Proposed Model

The model we propose [CG08] integrates the dynamical solution proposed by Bayerl et al [BN04] with a modified version of the vector averaging idea to provide both spatial preci-



(a) (b)

FIGURE 2.3 – The facilitation in four different MT neurons at 5 °, 6 °, 11 ° and 12 ° of eccentricity. The arrow indicates the preferred orientation of the neuron. Facilitation (red) effect occurs when the previous stimuli is in the direction of the preferred orientation of the neuron and suppression when is in the opposite direction (from Livingstone 2001).

sion and unambiguous motion perception. Our main contribution is to show that the dynamic propagation idea of [BN04] can be improved by a bio-inspired competition that allows to locally maintain the preeminence of the most activated filters while strengthening the detectors that correspond to the true velocity. Thanks to this principle, a coherent motion perception spreads along the edges of the moving object, with a better tolerance to noise.

The presented model is a sequence of two neural layers Ω_1 and Ω_2, see Figure 2.4(a). The first layer with an activity $p(\vec{x}, \vec{v})$ has as input the local motion information ($Input$), where the output of this layer is the input for the second one. In terms of the total number of neurons: $|\Omega_1| = |\Omega_2|$, but the second layer "sees" at the same time several units from the first one, see Figure 2.4(a). The second layer has an activity $q_1(\vec{x}, \vec{v})$ and it combines the response of several detectors in function of the spatial distance and the relative orientation among detectors, see Figure 2.4(b).

Each layer contains neurons that are associated to the components of vector $\vec{v} = (\Delta x, \Delta y)$, for example $(\Delta x = 1, \Delta y = 1)$ represents a movement of one pixel per frame to the right and one pixel upward. The variables $p(\vec{x}, \vec{v})$ and $q(\vec{x}, \vec{v})$ stand for the potential of the neuron at position $\vec{x}$ associated to $\vec{v}$ in layer Ω_1, respectively Ω_2. We associate the first layer to neurons in V1, and the second layer to neurons in area MT, with wider receptive fields and solving the aperture problem.

In order to avoid the divergence of the system, we include a normalization step in the second layer as proposed by [SH98], that we call $q_2(\vec{x}, \vec{v})$. The difference of the two levels is that the second layer performs a local competition in space and in the velocity domain, and the first layer receives feedback from the second layer, as first proposed by [BN04]. At a given time t, several iterations are performed keeping the same input: the successive iterations within the defined time t are indicated by the exponent n in our notation. The update state equation for the first level (Ω_1) is defined as:

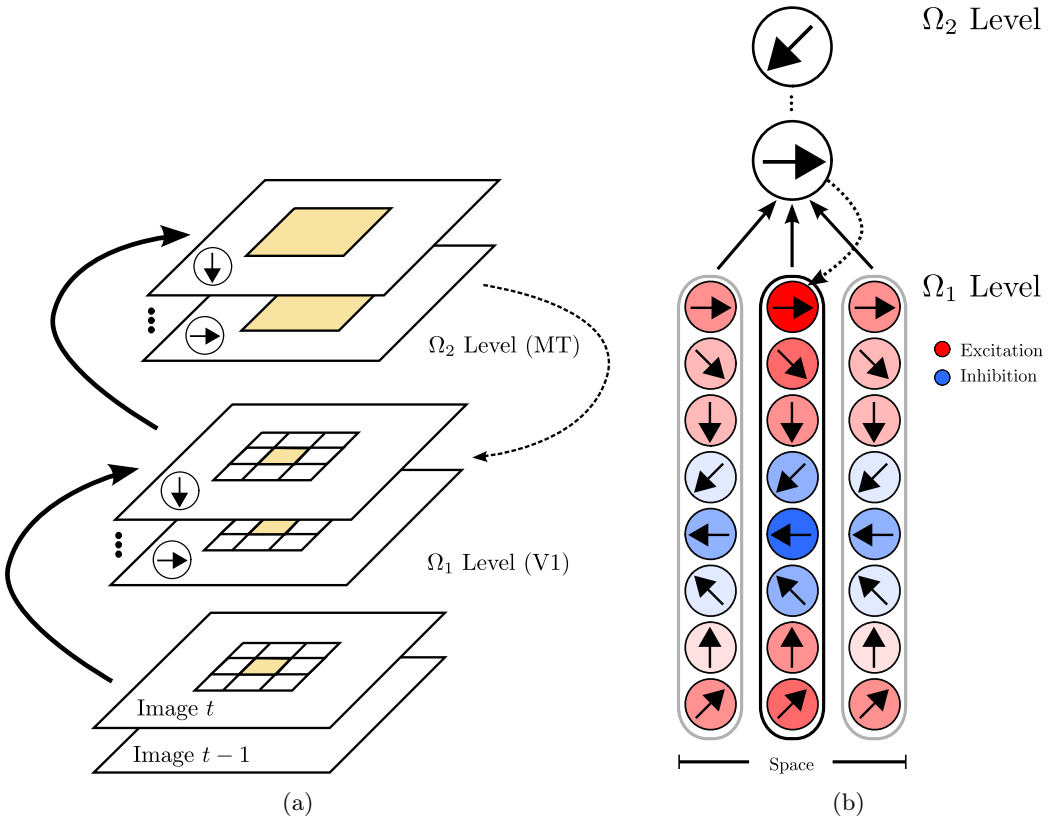


FIGURE 2.4 – Continuous lines are feed-forward connections and the dotted line is a 1-to-1 feedback connection. (a) General architecture of our model. There are several motion detectors at each pixel, as many as velocities at layers Ω_1 and Ω_2. The second layer has a larger receptive field than the first one. (b) Detail of the connections of Ω_1 and Ω_2. Each unit in the second layer receives positive influence from similar orientations (red) and inhibition from opposite orientations (blue), while the spatial distance modulates the interaction.

$$p^{n+1}(\vec{x},\vec{v}) \quad = \quad Input(\vec{x},\vec{v},t)\,(1+cq_2^n(\vec{x},\vec{v})) \tag{2.1}$$

where the input of the system is a local motion detection. The local motion detection must deliver, when the aperture problem occurs, a similar activity to all possible velocities. For unambiguous movements it must deliver a precise response. The parameter c is a control term for the level of feedback and the constant 0.01 avoids divergence of the system. The equation for the second level (Ω_2) is:

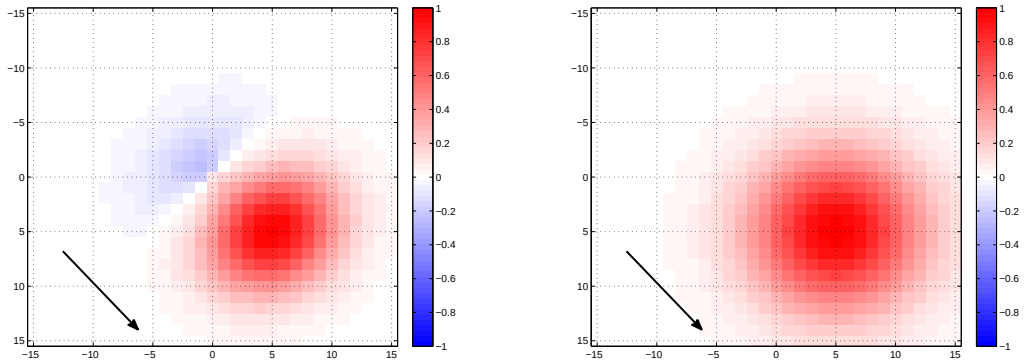
$$q_1^{n+1}(\vec{x},\vec{v}) \quad = \quad \left[p^{n+1} *_x G_{\sigma_1} *_v C_{\sigma_2}\right](\vec{x}',\vec{v}') \tag{2.2}$$

$$q_2^{n+1}(\vec{x},\vec{v}) \quad = \quad \frac{q_1^{n+1}(\vec{x},\vec{v})}{0.01+\sum_k q_1^{n+1}(\vec{x},\vec{v}_k)} \tag{2.3}$$

here the operator $*$ stands for the convolution, either in space ($*_x$) or velocity ($*_v$) domains. G_{σ_1} is a Gaussian function of parameter σ_1 and C_{σ_2} is our proposed kernel function that integrates different speed detectors (see below). The normalization is performed taking into account all the detectors at the same spatial location ($\vec{v}_k$).

The shape of the kernel C_{σ_2} is the main difference of our model with respect to the model of [BN04]. We have considered the evidence found by [LPB01], where opposite orientations within the MT receptive field have an inhibitory influence. We define the kernel in Eq. 2.4, using the relative angle of two velocities in the convolution that we note $\angle\vec{v},\vec{v}'$.

$$C_{\sigma_2}(\vec{v},\vec{v}') \quad = \quad G_{\sigma_2}(\vec{v}-\vec{v}')G_{\sigma_2}(|\vec{v}|-|\vec{v}'|)\cos\left(\angle\vec{v},\vec{v}'\right) \tag{2.4}$$



(a) (b)

FIGURE 2.5 – Kernels on the velocity space: (a) in our proposed model and (b) the model proposed by [BN04]. Both cases for an arbitrary speed ($\vec{v}=(5,-5)$).

We notice that our proposed kernel, see Figure 2.5(a), has a negative region and that the absolute value of the positive region is higher than the values in the negative region. This accounts for the idea that the inhibition is less strong than the coherent activation, thus supporting the soft inhibition idea discussed in [LPB01]. Compared with the receptive field structure depicted

in Figure 2.3, our kernel looks "reversed". However, Livinstone et al [LPB01] show that if the relative position of a moving target (one time instant before) is on the same direction as the oriented speed detector, hence there is a facilitation effect. Reversely if the target is located in the opposite sense. This idea can be then interpreted as a facilitation from similar orientation detectors, and inhibition from opposite orientations in the velocity space. The proposed kernel contrasts with the symmetric kernel proposed by [BN04] and other authors, see Figure 2.5(b), where the interaction (kernel) is always positive and only depends on the distance in the velocity domain (similarity of two speeds).

To summarize our model, in the locations of the image where the aperture problem appears, several motion detectors are simultaneously activated. In the first iteration the activity on the second layer is zero, thus there is no influence on the first layer. The second layer performs an "average" using our proposed kernel in a large area, where some detectors may be ambiguous and other unambiguous: the solution is to converge to the unambiguous response. The inhibition (negative values) we incorporate in our kernel, makes the model quickly discard "impossible solutions", i.e. to say left when the movement is to the right. It should be noticed that the error due to the aperture can be up to $\pi/2$ in right angles. After the second layer performs the local spatial and velocity convolution, it sends a multiplicative (positive) feedback. The feedback is multiplicative to only modulate regions where motion has already been detected (or static edges). Finally, as the system is recursive the process continues until the visual field (or the image) converges to a set of coherent movement regions. In the following section we discuss the empirical results of the proposed model both in synthetic and real sequences to evaluate our solution.

2.3 Results

The criterion we use to measure the error in the movement detection is the difference of the obtained motion direction and the true motion. The specific motion filtering we use is the ERD (Elaborated Reichardt Detector), a simple local motion detector based on the correlation of several oriented filters. We apply the filter proposed by [BN04] (see the Appendix of that work) to compare more precisely our results with their similar architecture. However, we introduce two modifications in their filtering method: we ensure positive outputs from the ERD (a negative activity means an opposite direction movement) and we compute the correlation in a 7×7 window, because in our images movements are relatively small. The correlation window detects horizontal movement of up-to 3 pixels-per-frame. We show the behavior of this filtering for a horizontal bar moving at speed $(-1, 0)$ in Figure 2.6. In all the following experiments we take the values $C = 100$, $\sigma_1 = 7$ and $\sigma_2 = 0.75$.

As we can notice in Figure 2.6, in ambiguous zones the ERD filtering generates multiple activations at the same level. The read-out of the velocity population is performed using a weighted average. For the depicted bar movement, at the corners the weighted average result is $(-1, 0)$ and in the middle location $(0, 0)$, even though several detectors are simultaneously activated. When the detected angle is 0 and the expected angle is different, we increase the error by $\pi/2$. In the following set of experiments we measure the angular error in the detection of our model that we call asymmetric excitatory-inhibitory (AEI) and we compare it against the method proposed by [BN04], that we call symmetric excitatory (SE).

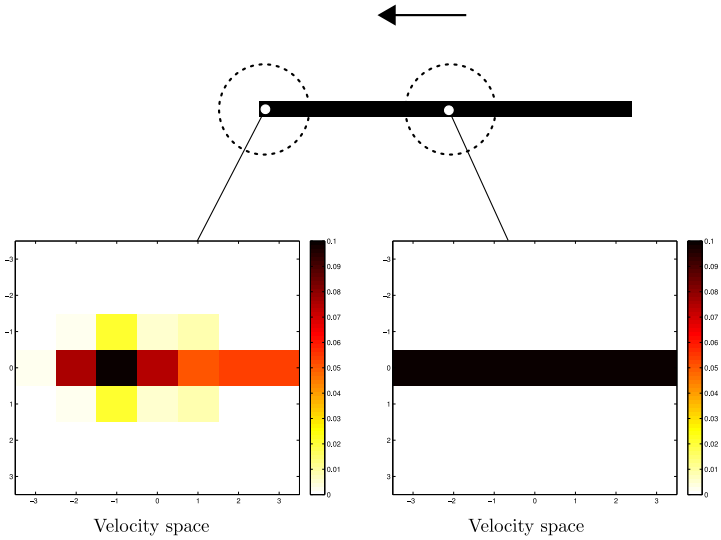
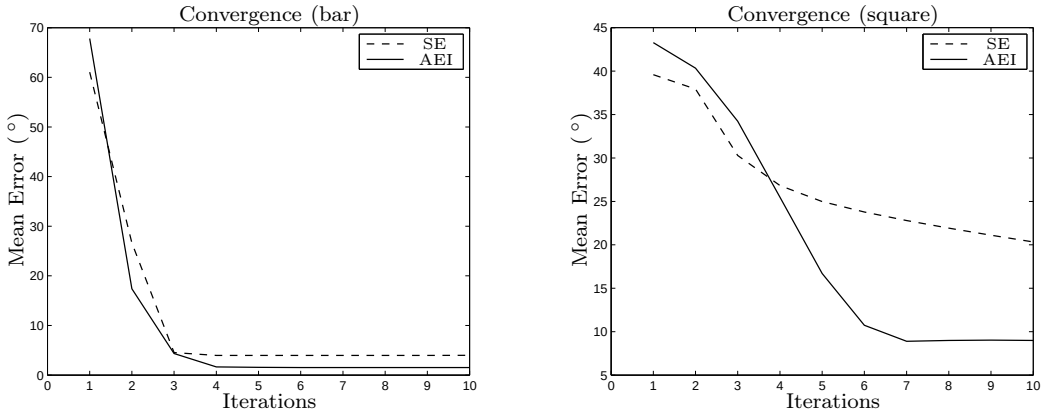


FIGURE 2.6 – The behavior of the ERD detector for a moving bar (arrow), when the aperture problem is not present (bottom left) and when the aperture occurs (bottom right). Note that the left detector has a larger activity at $\vec{v} = (-1, 0)$, which is correct, and the detector in the middle of the bar responds with the same amplitude to $\vec{v} = (a, 0)$, where $a \in \{-3, -2, -1, 0, 1, 2, 3\}$.



(a) (b)

FIGURE 2.7 – Convergence of our proposed model (AEI) compared with the symmetric excitatory only model (SE) for the cases of a moving bar (shown in Figures 2.8(a), 2.8(b), 2.8(c)) and square (Figures 2.8(d), 2.8(e), 2.8(f)). For the two models the convergence of the layer interpreted as V1 was taken into account ($p(\vec{x}, \vec{v})$ in our model).

2.3.1 Moving bar

Figure 2.8 (first row) shows our simulation for a horizontal bar moving in the horizontal direction, with a velocity $\vec{v} = (-1, 0)$. Detection is correct at the ends of the bar before starting (iteration 0) and the aperture problem appears along the bar. Our results show how many detectors give the right direction as a function of the number of iterations, see Figure 2.7(a) for both AEI and SE model. The detection is correct with the two models and the convergence time is not significantly different. Probably, this similarity in the convergence time is due to the fact that the system interacts mostly within the positive part of our proposed kernel and by consequence there is no important difference among the models.

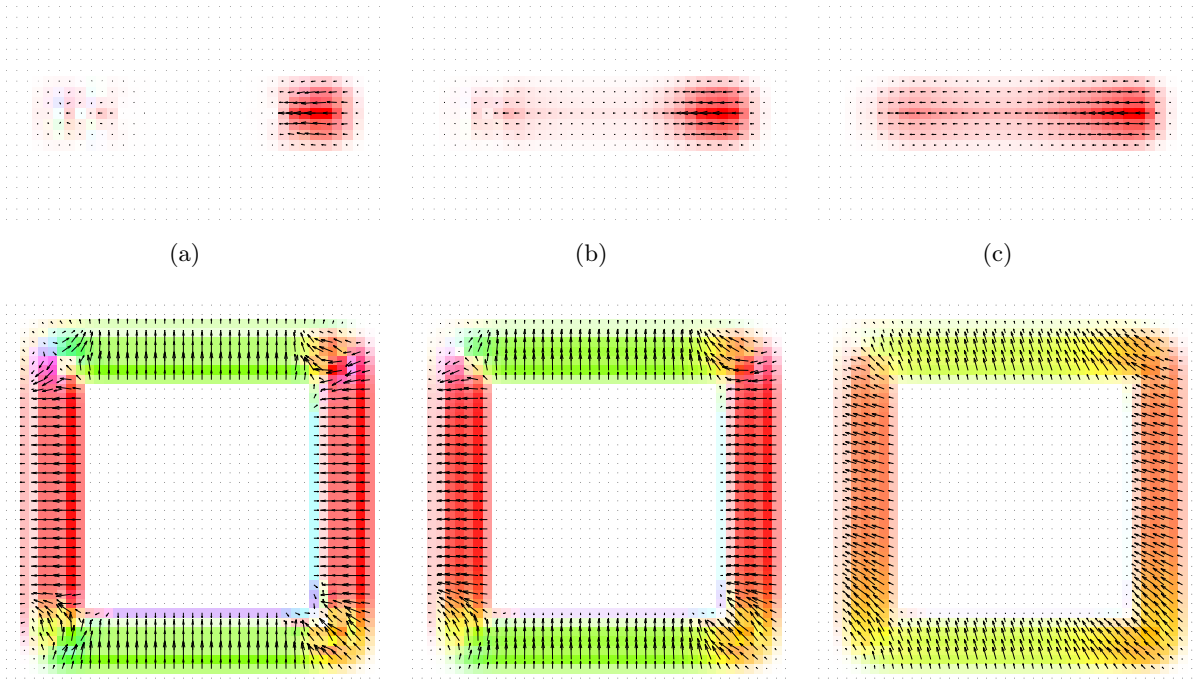
2.3.2 Diagonal moving square

The second row of Figure 2.8 shows our experiment with a solid untextured square that moves diagonally, with a velocity $\vec{v} = (-1, 1)$ in pixels per frame. Detection is correct close to the corners from the first iteration as expected (we draw the orientation with different colors and speed with different intensities), but along the borders the aperture problem can be observed and wrong directions appear because of the discrete nature of the object, see Figure 2.8(d). Our results show how many detectors give the right direction as a function of the number of iterations, see Figure 2.7(b), and the effect of lateral inhibition (AEI) when we compare with the SE interaction. The diagonal square experiment shows that the convergence is significantly faster with our AEI kernel than with the SE one, see Figure 2.7(b). Probably this difference is due to the fact that more detectors in the velocity space are activated in this case, giving a more important role to the negative part (even though of small amplitude) of our proposed kernel (inhibition).

2.3.3 Real sequences

The next experiment was to study real sequences, where the ground-truth optical flow was available to compare directly with our proposed method (AEI). We make use of a public domain database from [MNCG01], where the movement of real, yet simple, objects is available with its ground truth displacement. From this database, we use two of their sequences: a translating checkboard and a camera doing a zoom centered onto a box, see Figure 2.9. These sequences are interesting because the initial movement detection is noisy and the background is textured, with many wrongly activated movement detectors. At the same time, the flow is not regular and it has discontinuities in both cases.

The results of our model for the real sequences can be seen in Figure 2.10. In the checkboard sequence, where a very strong noise can be noticed because of the textured background and compression effects, our model shows a better tolerance to noise, however it converges slightly slower than the SE model. In the case of the optical flow of the zoom sequence, it is more subtle: as expected the optical flow is expansive (because of the zoom), but the box in the middle generates a small discontinuity in the flow, as shown in the ground truth signal of the zoom sequence in Figure 2.10. In this more complex scenario, the convergence speed of the two models is similar, but our model better preserves the discontinuities in the flow, despite the small difference in amplitude (close to the box). The SE model gives much smoother results, almost completely removing discontinuities. When compared with the SE model, our model introduces small perturbations in the amplitude of the velocity (intensity of the color) but it better preserves the discontinuities.



(a) (b) (c)

(d) (e) (f)

FIGURE 2.8 – Motion detection evolution for the Ω_1 layer (or $p(\vec{x}, \vec{v})$) that represents V1. Upper row shows motion detection for a moving bar for (left to right) an: initial detection, after 2 iterations and after 4 iterations. Here the red color represents true movement. The bottom row shows a similar analysis for a moving square. The dark yellow color marks here the true motion.

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

FIGURE 2.9 – Two frames from the checkboard and the zoom sequences from the ICCV DB.

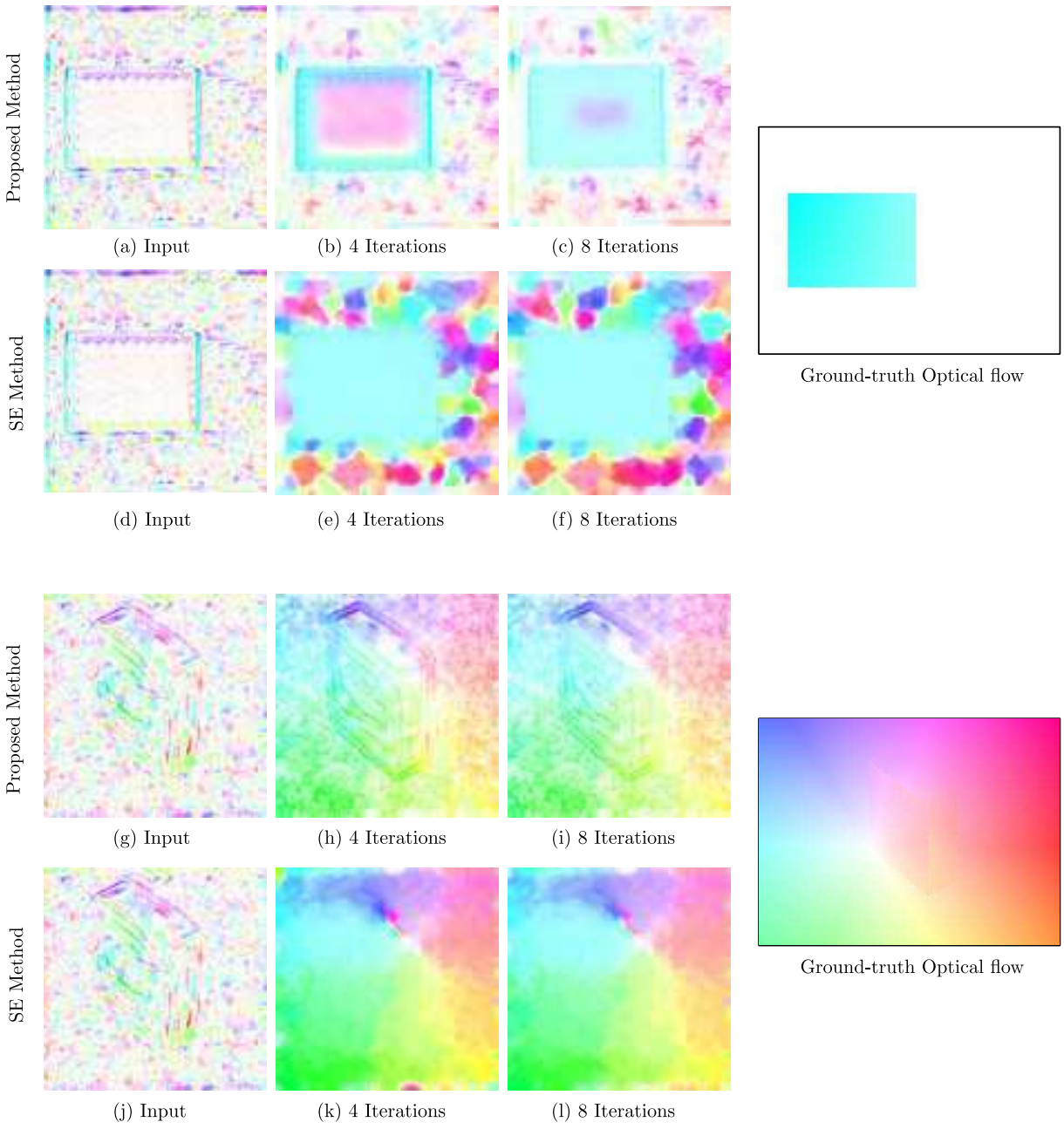


FIGURE 2.10 – The evolution of the speed detection for two sequences: a translating checkboard (first two rows) and a zoom over a box (last two rows). For the two sequences we use our AEI model (first and third row) and for comparison the SE model (second and fourth row). We use 0, 4 and 8 iterations (left to right) in all illustrations. The two panels in the last column show the true motion signal. The optical flow is in the Middlebury color code [BRS+07] (color represents orientation and saturation represents speed).

2.4 Discussion

Many models inspired by brain circuitry have been proposed for the local detection of motion [WA95, SH98, BN07b, CH10], proposing hypotheses about how real neurons may interact in terms of functionality. All these models deal with the aperture problem with different degrees of success. Despite this, we note that the aperture problem never arises alone, as the local detection of movement is subject to other types of perturbation: false positive activation, non-binary responses (higher responses at correct velocities but non-zero at other velocities) and contrast variances among others. In our work we propose a model that mainly handles false positive activations of local motion detectors, solving the aperture problem, including in noisy scenarios. Our model achieves this solution without losing spatial precision by incorporating local inhibition in the kernel for inconsistent movement detection (negative values). Using computer simulations, our experiments show that this lateral inhibition improves the results when noise is present and improves or maintains the convergence speed in most cases when we compare it with excitatory-only models.

Two reported elements of the primate brain that can be retrieved in our model are the local nature of the motion detection and a convergence time [OSK03, BFM10]. At the same time, we propose an interpretation to the microcircuitry found in area MT, where facilitatory and inhibitory effects have been reported for speed detection neurons. Our interpretation is that this mechanism improves the convergence of the system and at the same time makes it more robust in noisy scenarios. We note, however, that these facilitatory/inhibitory effects are dynamic and not static (the kernel changes its shape), thus our model presents a first interpretation of the functional nature of this neural mechanism. The dynamics of these interactions may be associated with an initial denoising process that is attenuated over time.

The proposed model for the detection of movement, using local information, presents a distributed computation where the dynamics of a population of units (neurons) is able to: (1) propagate the information in space independently of the size of the moving object and also of the receptive field size of each neuron, and (2) deliver these results under noisy conditions where local movement detectors may be wrongly activated. These two ideas allow building a robust distributed mechanism for the solution of the aperture problem and in some cases a reduction in the convergence time. The architecture we present illustrates how a population of local motion detectors can deliver an effective solution to a problem that cannot be solved locally through the dynamics of the population. More precisely, in our model we highlight that in noisy scenarios better results can be obtained by including the interaction within the population, not only as a spatial/velocity average, but also interpreting the activity of each unit with respect to neighbors units in the population.

3
Speed Sampling

Contents

Most optical flow extraction techniques work within a finite range of speeds. Usually, the range of detection is extended to higher speeds by combining some multi-scale information in a serial architecture. This serial multi-scale approach suffers from the problem of error propagation related to the number of scales used in the algorithm. Biological experiments show that human motion perception seems to follow a parallel multi-scale scheme. In this chapter we present a bio-inspired parallel architecture to perform local detection of motion, providing a wide range of operation and avoiding the error propagation that is associated with serial architecture. To test our algorithm, we performed relative error comparisons between both classical and proposed techniques. This shows that the parallel architecture is able to achieve motion detection with results similar to the serial approach usually implemented in computer vision. We also show that our algorithm retrieves perceptive properties found in human vision as the relative error in speed detection. At the end of the chapter we discuss the plausibility of our model and explain how it provides a distributed mechanism to increase the range of speed detection.

3.1 Speed coding

In this work we are interested in the local detection of motion, specifically in the coding and retrieval of speed ($\vec{v}$), and in the link between the idea of selecting a range of speed to work with, and providing wider ranges of discrimination, as observed in human psychophysical experiments [OCDBM85]. We focus on two features: the multi-scale architecture of speed detection, and the

relation between the number of multi-scale levels and the range of speeds the system is sensitive to.

3.1.1 Motion detection in computer vision

The detection of motion is a widely used operation in computer vision. Commonly called "optical flow extraction", the main objective is to assign a vector $\vec{v} = (u, v)$ to each frame pixel from a given sequence of frames, see section 1.1.2 for more details. In this section, we explain the basic technique to increase the motion range of an optical flow extraction to which the method is sensitive. We ground our explanation on the well-known Lucas & Kanade's method [Luc85, BFBB94] (the basic multi-scale technique similarly applies to other methods for optical flow extraction).

3.1.2 Serial multi-scale optical flow

The Lucas & Kanade method for optical flow extraction considers a small region of the visual field (Ω). The use of this region Ω is not particular to this method: it is used in most algorithms [BFBB94]. As the computation is performed in small windows, the detection of motion is constrained to detect speeds up to ω pixels per frame, where ω stands for the diameter of Ω. To overcome this limitation, a multi-scale representation of the images can be performed, usually by using Gaussian pyramids [Bla92]. A Gaussian pyramid representation of an image is computed by recursively smoothing (using a Gaussian kernel) and sub-sampling the original image. In this way, the original image is represented by a set of smaller images. The representation at scale level $l = 0$ is the original image itself. The image at level l is obtained by sub-sampling a filtered version of the image at level $l - 1$ with a downsampling factor equal to 2. Thus, the size of the image at each level l is $N_l = N_{l-1}/2$ with $l = 1, 2, \ldots, (L - 1)$, where L is the number of levels of the representation.

In the serial multi-scale optical flow estimation, speed is computed by sequentially projecting the estimation obtained at level l to level $l - 1$, until level $l = 0$. But there are more complex strategies for computing the optical flow with a multi-scale approach [Sim99]. We use the Lucas & Kanade's algorithm [Luc85, BFBB94] taking advantage that it is implemented in the widely used computer vision library OpenCV [BK08, Bla92]. In this case, the multi-scale estimation starts from the highest level ($l = L - 1$) and it propagates to the next one:

$$\vec{v}_{l-1} = 2\vec{v}_l + d_{l-1}(\vec{v}_l) \tag{3.1}$$

where d_{l-1} is the estimation of velocity at level $l-1$ after projecting the estimation v_l by warping the image by $-v_l$ at level $l - 1$. Since at the level $l - 1$ there are more pixels than at level l, the speed estimated at a higher level must be interpolated into the lower level. Accordingly, we interpolate using the closest four points in our implementation. Computing the optical flow from the highest level and then projecting the solution to the lower level [Sim99, Bla92] increases the range of detectable speeds. This range is wider when more scales are used. The sequential projection among levels also propagates the error introduced at each level. Thus, in terms of absolute precision, increasing the number of scales in the representation increases the error of the estimation.

3.1.3 Biological elements

This section sketches out the current experimental knowledge in biology, focusing on studies of speed coding in the human brain [NHD05] and on higher level descriptions of speed discrimination from experimental psychophysics [MN84, Kvv85, MVB94].

Parallel architecture

In the human brain the main area that is responsible for coding different speeds is area MT [SZ03]. It is located in the occipital region (back of the head). Neurons in this area are mostly selective to stimuli moving at a given speed [MVE85]. Their spatial organization is retinotopical [LN03]: each neuron has a reduced visual field, and neurons which share the same local visual field are grouped together in cortical columns that are selective for different orientations, see section 1.2 for more details. This configuration allows a complete mapping of the visual field with a group of cortical columns that encode for all possible directions of local motions. The spatial organization is less known with respect to the speed selectivity. Nevertheless, it has been found that (1) the average detected speed increases with eccentricity (with respect to the retinotopical organization of MT), (2) similar speeds are detected by neurons closer than for distinct speeds, and (3) for each eccentricity, there are neurons for different speeds [LN03]. The interactions among different units are not completely understood, but there is evidence that units sensitive to different speeds could be coding a range of speeds in parallel [MVB94]. It has also been observed that the range of detectable speeds is not uniformly covered [MVE85]. Considering these elements, in this work we are interested in how speed selective units in MT can deal with the estimation of speed working in a parallel architecture and showing human-like discrimination capabilities.

Speed discrimination

In the work of McKee et al. [MN84], two subjects were exposed to several stimuli, one of those being a horizontal scaled single bar vertically moving at different eccentricities. The goal of this experiment was to determine the minimal relative detectable variation in speed for every subject with the sight fixed at a certain location and for each stimuli eccentricity.

It is important to mention how this was actually measured, because the subject cannot assign a precise velocity at each location. Instead, given a reference velocity, the subject was asked to indicate whether the next presented stimuli moves faster or slower. The minimal detectable variation was then statistically inferred. Related experiments were performed by others [OCDBM85, MVB94, Kvv85], showing that the measurements are not affected by different contrast conditions, and that they do not depend on binocular or monocular sight.

The described experiments study speed discrimination at several eccentricities [4], see Fig. 3.1. We point out that in these experiments only the left side of the experimental curves in Fig. 3.1 is related to the eccentricity, so we can argue that for each eccentricity there is a wide range of speed discrimination where the relative error (rather than the absolute error) remains stable (5%-15%). In this chapter we are interested in speed discrimination at one eccentricity (any of them) and its associated discrimination properties, and not in the relations among different eccentricities. To model these discrimination functions, we need to generate a given discrimination percentage (related to relative error, 5%-15%) in a given range of speed $[v_1, v_2]$.

4. Distance to the center of the eye in foveated vision (humans, primates and others).

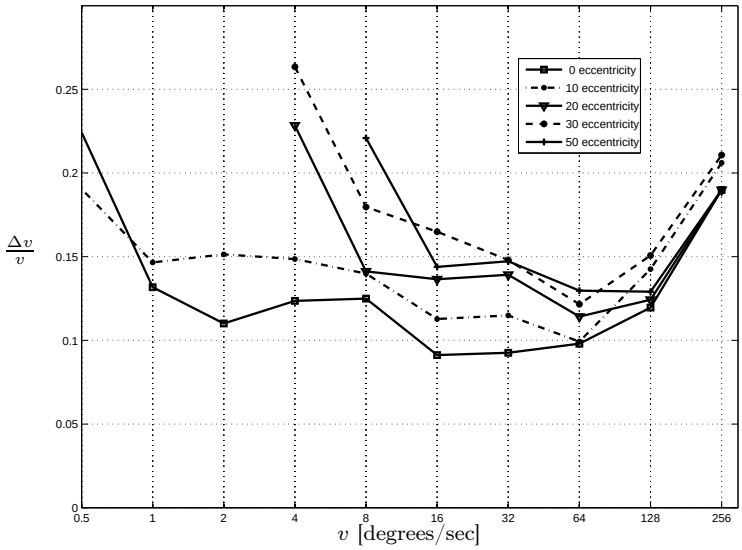
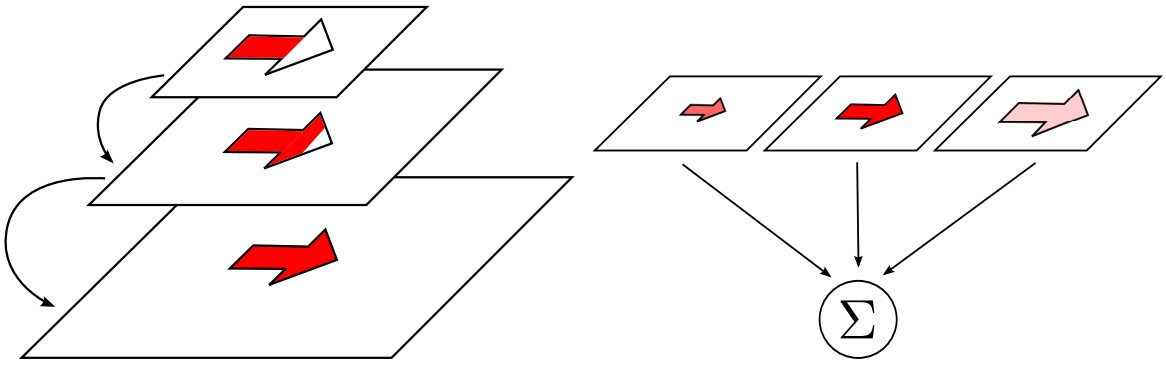


FIGURE 3.1 – Weber fraction (minimum threshold of perceived change in speed or $\Delta v/v$) as function of velocity for different eccentricities in the human subject B.D.B. (condition 2). In the differential motion experiment performed by [OCDBM85] (the data is reproduced from their work), each curve is at eccentricities $0\,^\circ$, $10\,^\circ$, $20\,^\circ$, $30\,^\circ$ and $50\,^\circ$, respectively (left to right). The speed axis is in logarithmic scale.

3.2 Proposed parallel multi-scale speed detection

Multi-scale speed detection is based on the fact that a particular speed detection algorithm can be used to estimate slower speeds at lower levels and to estimate faster speeds at higher levels. This information is used in the above described serial multi-scale optical flow algorithm to detect speeds in a wide range of velocities; by first computing high speeds (in a higher level), then using that estimation in lower level as an initial solution and performing a new estimation. In general, the same process can be repeated projecting the information at level $l+1$ to estimate speed at level l, *i.e.* in a serial manner, see Figure 3.2(a). As it is described in [MVE85, NHD05], it seems that the human motion perception is based on a parallel multi-scale scheme. Based on this idea, we propose a speed detection algorithm that estimates the speed by combining the information computed at each level independently, *i.e.* using the multi-scale information in a parallel manner, see Figure 3.2(b). In this case, there is no error propagation on the computation of speeds at each level because it does not depend on the estimation performed for other levels. At each level l, we compute speeds using the optical flow estimation algorithm described above, see subsections 3.1.2 and 1.1.2. As explained before, our choice does not bias our results, since the objective is to provide bio-inspired parallel speed detection instead of the standard serial approach for any optical flow extraction method.

When speed has multiple estimations, a certain confidence must be assigned to each one of the estimations to combine them. To define this confidence, we notice that any speed detection algorithm estimates speed with a certain relative error. While the absolute error is the difference between real an estimated speed or $|v_r - v_e|$, the relative error is the absolute error with respect



(a) (b)

FIGURE 3.2 – (a) In the serial approach confidence, in red, increases as estimation is propagated. (b) The parallel approach simultaneously estimates several speeds at different confidence level.

to the real speed or $\left|\frac{v_r - v_e}{v_r}\right|$. Taking these definitions, we propose a simple confidence measure $k_l(v_r)$ for the estimation of speed at each scale. It is,

$$k_l(v_r) = 1 - \left|\frac{v_r - v_e}{v_r}\right| \tag{3.2}$$

where v_r is the magnitude of the object's real speed ($v_r = \|\vec{v}_r\|$) and v_e is the magnitude of the average estimated speed on the object pixels location. When the relative error is zero (or close to zero), the confidence is maximal ($= 1$). It can be noted that this computation only takes into account the magnitude of the speed, ignoring its direction. Figure 3.3(a) shows the confidence $k_l(v_r)$ for three different multi-scale levels. These distributions were computed using an input image sequence containing an object moving at different speeds in a range from 0.5 pixels per frame to 20 pixels per frame. To statistically determine the confidence at each level l and speed v_r, experiments were carried out using the input image sequence with several realizations of Gaussian white noise, then the resulting confidence $k_l(v_r)$ was computed as the mean value of the ones obtained in the experiments. Figure 3.4(a) shows a frame of an input image sequence used in the experiments. In this sequence the object is moving at 10 frames per pixel in the bottom-right direction with Gaussian noise.

As it may be seen in Figure 3.3(a), a particular speed v_r can be detected at several multi-scale levels but with different confidence values. Thus, the current speed could be estimated by taking into account the speeds computed at each level l and their associated confidence values k_l. For that reason, the experimental distributions depicted in Fig. 3.3(a) have to be approximated by a closed-form equation. We model these distributions as Gaussian distributions in a semi-log space defined by the following equation,

$$\hat{k}_l(v_r) = \exp\left(-\left[\frac{\log(v_r) - \mu_l}{\sigma_l}\right]^2\right) \tag{3.3}$$

$$\mu_l = \mu_0 + \log(c^{l-1}) \tag{3.4}$$

$$\sigma_l = \sigma_0 \tag{3.5}$$

where μ_0 and σ_0 are the mean and variance of the distribution at level $l = 0$ and c is the

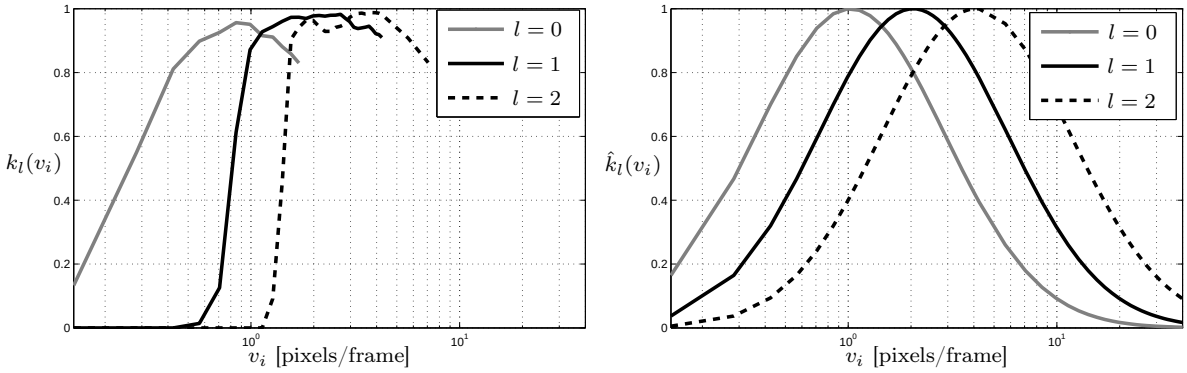


FIGURE 3.3 – Confidence distribution k_l for different levels l. (a) Experimental distributions k_l. (b) Approximated distributions $\hat{k}_l$.

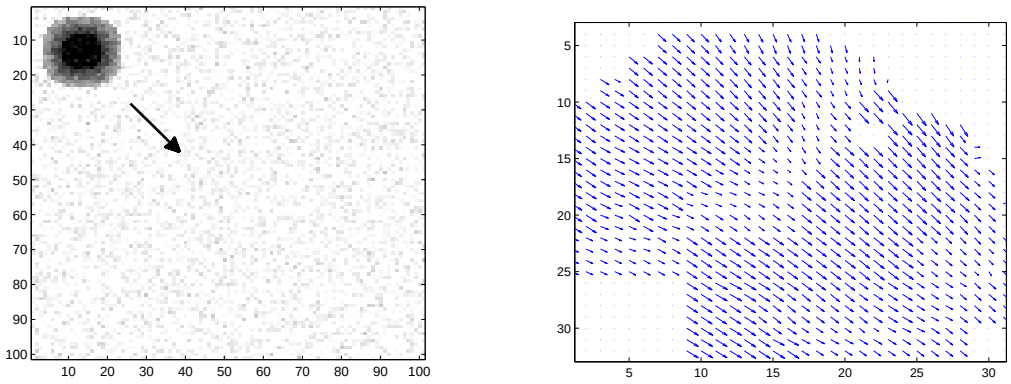


FIGURE 3.4 – (a) One frame of an input image sequence used in the experiments is depicted ($v = 10$ bottom-right direction). (b) Obtained optical flow using the proposed parallel multi-scale algorithm. Note that optical flow image was zoomed around the object position.

scaling factor used in the sub-sampling of the images. The approximated distributions $\hat{k}_l$ for $l = 0$, $l = 1$ and $l = 2$ are depicted in Fig. 3.3(b). A better approximation of the distributions could be obtained using a particular set of variables for each level but this would increase the model complexity. The approximation of the distributions $\hat{k}_l$ for each level in Eq. (3.3) only depends on μ_0, σ_0 and c. It is noteworthy that our approximation allows to perform the estimation of speeds using different values of the scaling factor c, which is usually set to $c = 2$, as in the case of using Gaussian pyramids for the sub-sampling.

Finally, denoting the detected speed at each level l by $\vec{v}_e^{\,l}$, the proposed algorithm computes the current speed using the speed detected at each multi-scale level with its associated confidence value $\hat{k}_l(\|\vec{v}_e^{\,l}\|)$, as

$$\vec{v}_f = \frac{\sum_{l=0}^{L-1} \vec{v}_e^{\,l} \hat{k}_l(\|\vec{v}_e^{\,l}\|)}{\sum_{l=0}^{L-1} \hat{k}_l(\|\vec{v}_e^{\,l}\|)} \tag{3.6}$$

where L is the number of levels used to compute the estimated speed $\vec{v}_f$. Figure 3.4(b) shows the obtained optical flow using the proposed parallel multi-scale algorithm. The comparison between the experimental confidence distribution of the proposed algorithm and confidence distributions for four levels is shown in Fig. 3.5. As expected, the confidence distribution of the parallel multi-scale algorithm with $L = 4$ is approximately the envelope of the confidence distributions of levels $l = 0$, $l = 1$, $l = 2$ and $l = 3$

In contrast, the confidence of the serial optical flow estimation, see Figure 3.6, shows sudden drops. This effect is due to error propagation (aliasing), and implies that the confidence decreases as more scales are used the. Taken separately, each scale has a similar maximal confidence level (see Figure 3.5), but serially combined errors can be projected onto lower scales. Once a lower scale receives a wrong estimation, it cannot recover from these errors and the error will start to be accumulated. This problem is more evident when the same speed can be "seen" by several scales at the same time, i.e. when curves overlap in Figure 3.6.

3.3 Results

As it was described in subsection 3.1.3, speed discrimination is inferred as the minimal detectable variation in speed of a particular visual stimuli in psychophysical experiments on humans. In this section we aim at verifying that our proposed parallel architecture can reproduce these results, and that the classic serial multi-scale estimation cannot, because of its error propagation. First we define speed discrimination computationally (noticeable speed variations) in our context, where we have a precise value for $\vec{v}$. A variation of speed, from a given reference speed v_{obj} of the moving object, is defined noticeable if the following inequality holds,

$$\left| \frac{\widehat{v}_{v_{obj}} - \widehat{v}_{v_{obj} \pm \Delta v_{obj}}}{v_{obj}} \right| > \alpha \tag{3.7}$$

where $\widehat{v}_{v_{obj}}$ and $\widehat{v}_{v_{obj} \pm \Delta v_{obj}}$ are the speeds estimated by the algorithm when the object is moving at velocities v_{obj} and $v_{obj} \pm \Delta v_{obj}$, respectively, and α is the percentage of variation (from the object speed v_{obj}) required to select Δv_{obj} as detectable. It may be noted in Eq. 3.7 that a variation in speed is noticeable if it is detectable when v_{obj} is both increased and decreased by Δv_{obj}. To statistically determine the minimum value of Δv_{obj} several (100) experiments were carried out using the input image sequence with several realizations of a Gaussian white noise. Then, the minimal detectable variation in speed, from a given reference speed v_{obj}, was computed as the minimum detectable Δv_{obj} obtained in 90% of the experiments.

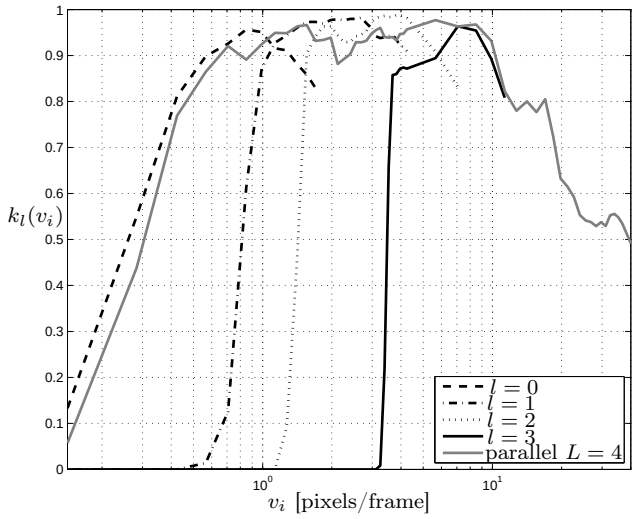


FIGURE 3.5 – Comparison between confidence distribution of the proposed algorithm (parallel) with $L = 4$ and $c = 2$, and confidence distributions k_l for levels $l = 0, 1, 2$ and 3 without projection between them.

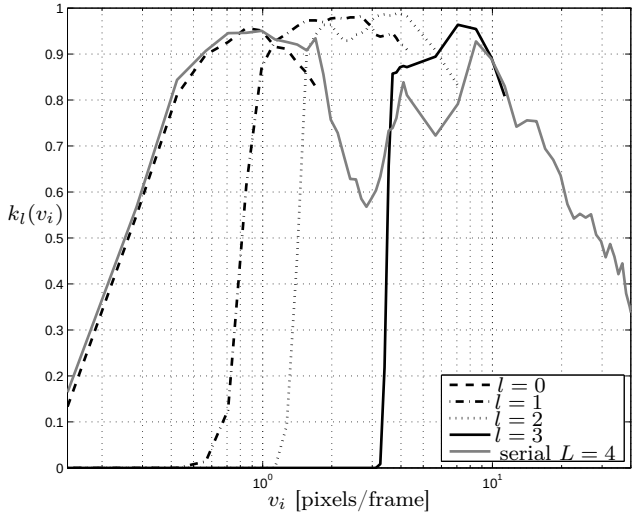


FIGURE 3.6 – Comparison between confidence distribution of the serial algorithm using a Gaussian pyramid with $L = 4$ and $c = 2$, and confidence distributions k_l for levels $l = 0, 1, 2$ and 3 without projection between them.

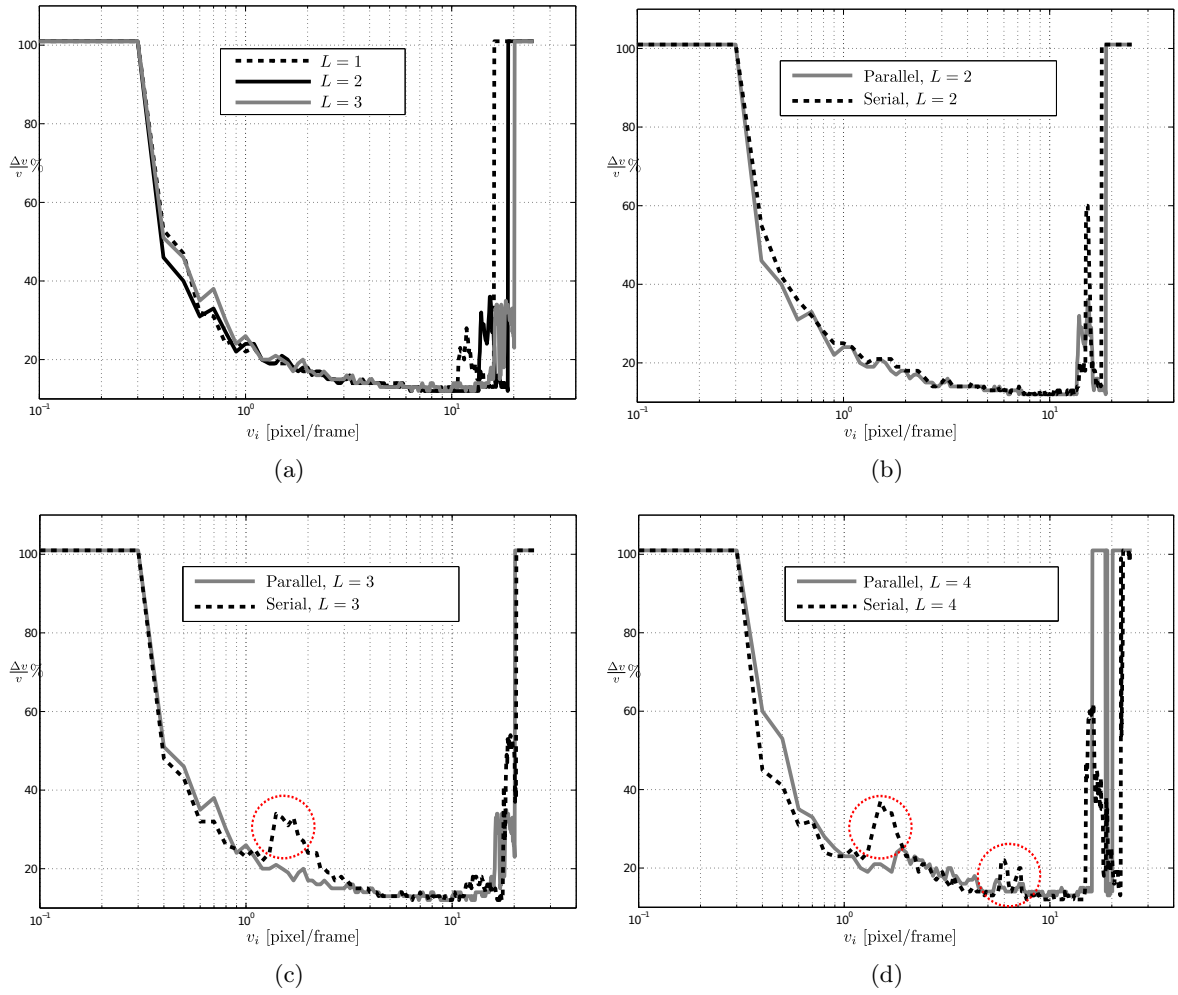


FIGURE 3.7 – (a) Discrimination of the proposed parallel algorithm for $L = 1$, $L = 2$ and $L = 3$. In (b), (c) and (d), the comparisons between the discrimination of the proposed parallel and the serial algorithms for $L = 2$, $L = 3$ and $L = 4$, respectively, are shown. In all the simulations, parallel and serial algorithms use c = 2.

We summarize our results in Fig. 3.7. Figure 3.7(a) shows the discrimination of the proposed parallel multi-scale algorithm for different values of L. The range of discriminated speeds is enlarged when the number of levels used in the multi-scale representation increases (right part of the curves). In comparison to the serial multi-scale, our method has a similar range of speed discrimination when the same number of levels are used, see Fig. 3.7(b), 3.7(c) and 3.7(d). When observing the mean and variance of the discrimination in the range of speeds from 1 to 15 pixels per frame, the parallel multi-scale method shows lower values, and then better discrimination. For the case of $L = 3$, our parallel discrimination method has mean= 14.1 and variance= 2.2, while serial discrimination has mean= 15.5 and variance= 4.2. It indicates that the proposed parallel algorithm presents a better discrimination in this range. It can also be noticed that for $L = 3$ and $L = 4$, respectively Figures 3.7(c) and 3.7(d), several bumps (see the red circles) in the serial algorithm are observed. They correspond to the zones of lower confidence in the serial algorithm. By consequence, the error propagation of the serial algorithm can be also found when discrimination is measured, while our proposed parallel implementation shows a regular behavior, closer to the observed speed discrimination in humans.

3.4 Discussion

Recent works [Sim99, CGM98] have developed the idea of a multi-scale estimation of speed. Simoncelli [Sim99] proposed a Bayesian scheme to compute the error distributions and then estimate velocity using a Kalman filter over the space of scales (not time). This approach builds a far more sophisticated error function, but is still serial. Our work assumes that the error functions are fixed, while Simoncelli [Sim99] assumes that the error changes with respect to $\nabla I(\vec{x}, t)$ because higher values of the gradient can be associated with locations where the aperture problem is absent. Chey et al. [CGM98] proposed that higher threshold levels for higher scales (scale-proportional thresholds) and inter-scale competition could explain human speed discrimination curves. We have presented a scheme where the response of each scale regulates the relevance of the responses of that scale. This regulation (confidence function) could correspond to the notion of threshold and the combination of responses to the idea of competition. To our knowledge, no other previous study models error functions as Gaussians in the log space. The proposed log-distribution of speed detectors combined by a vector average population read-out could explain human speed discrimination curves. Moreover, it also seems to fit recent recordings of distribution of motion sensitivity of neurons in area MT [NHD05].

The presented model takes inspiration from human physiology and psychophysics knowledge in the sense that it achieves wide uniform relative discrimination properties by using an evenly spaced set of detectors in the logarithmic space. In addition, the computation time is no longer a function of the number of scales, because the speed estimation on each scale is performed simultaneously. Our approach estimates the number of different motion detectors we should include (for each orientation) to have a certain speed detection range, where the response will emerge from the interaction of different local motion detectors (the weighted average). The model can be also understood as a possible population encoding of movement, where for a given speed, possibly several units will be activated (each with a preferred or optimal speed). The read-out of the speed detector population we propose can, at least, verify the homogeneous speed discrimination found in psychophysical experiments.

In terms of distributed computation of movement, in this chapter we explain how local movement detectors can be organized to obtain a certain global response (speed discrimination, in this case), to have human-like speed discrimination capabilities. Accordingly, the results can still

be compared to algorithms in computer vision, linking relative error and speed discrimination. Moreover, the parallel computation of speed we propose does not depend on a serial process making the speed estimation algorithm suitable for parallel implementations.

Taken together, the solution for the aperture problem in chapter 2, and now the sampling of speed present a general framework to compute velocity in the context of distribution computation, addressing two important aspects: the orientation and amplitude of movement. The two models illustrate how the perception of local movement can be understood as the result of the dynamics of a population of units, where the interactions among neurons deliver the answer of the system.

Part II

Cognitive vision

4

Features and discrimination

Contents

The first part of this thesis studies how optical flow extraction can be performed in a distributed framework. We have presented models about two aspects of this problem: the aperture and multi-scale speed detection, to solve them in a distributed architecture. The second part continues the idea of distributed computation, now in the context of motion pattern classification. The main objective is to study how this problem can be solved in a distributed framework, taking inspiration from biological systems and comparing these to computer vision techniques. However, as we will see in the chapter, pattern analysis is closely related to feature extraction, e.g. motion detection. Following the presentation of our pattern classification model in Chapter 5, the importance of feature extraction will be stressed in Chapter 6 by using different feature extraction techniques.

In this Chapter we will introduce the most common techniques in computer vision to perform feature extraction and human action recognition from video signals. At the same time, we make a general review about feature extraction in the human brain and the discrimination of spatiotemporal sequences. The spatiotemporal processing, the representation and subsequent sequence discrimination is not as extensively known to biologists as motion perception and they are currently subject to intense debate. The main objective of the chapter is to introduce and prepare the discussion for the following chapter, where our model of motion pattern classification is presented in detail.

4.1 Computer vision

In Chapter 1 we introduced an initial discussion about computer vision chronology. Here we briefly extend this idea to introduce current methods of feature extraction and motion pattern recognition.

Considering the latter part of the 20^{th} century, applications seem to be one of the major challenges currently in computer vision. To face this challenge, researchers have followed solutions from AI, pattern recognition, or more pragmatic approaches where engineering aspects and adaptation to specific problems are crucial for their solution. To analyze these approaches we discuss the Speechome project [DR06].

In the Speechome project, microphones and video cameras were installed in the rooms of a normal house, where a family of three adults and two children freely interacted. One of the goals of the project was to give insights about child language development. As part of this goal, thousands of hours of video and audio were analyzed, for example, to label different human actions as: "walking", "eating" and "playing". Despite the extreme nature of this project, it raises one key question from a computer vision perspective: can current methods in computer vision and pattern recognition deal with flows of data recorded in realistic conditions? It seems that the answer is no. This is what the authors of the Speechome project concluded [DR06] after six months of operation, subsequently choosing semi-automatic labeling techniques adapted to the problem. I also share this point of view from my experience in computer vision applications, where the adaptation to the problem is crucial [CHKM07].

Keeping in mind that applications are important to perform visual pattern recognition tasks, in this Chapter a review will be presented about feature extraction and pattern recognition techniques related to the human action recognition problem. First, some of the techniques that are commonly used for feature extraction and pattern recognition from videos applied to the human action recognition problem are introduced. In the second part, a review of current knowledge in biology will be presented to compare the two approaches.

4.1.1 Feature Extraction

Related to human action recognition, a huge list of techniques can be found [MG01, MHK06, Pop10]. In this work, we divide feature extraction methods into local or global techniques, depending if features require local or global information to be computed (in the spatiotemporal volume): in one extreme case using only one pixel information and in the other extreme case the complete silhouette of the human pose where almost all pixels are required. This classification becomes arbitrary when statistical or probabilistic approaches induce local operations in the spatiotemporal volume. In these cases we consider the technique global, as it requires a previous analysis using the complete spatiotemporal volume, in order to derive a local operator.

Local features

Local features process information only in a small neighborhood in the spatiotemporal volume. These techniques have been used in many implementations, as they are easy to parallelize/distribute, in order to compute them fast (in real time). We present a short list of features extraction techniques, based on local computations, to exemplify some of these methods.

Edge detection. Edge detection searches for the frontier among objects or zones with different light/texture/color, thus zones of change. To perform edge detection, we find techniques based on masks derived from mathematical operators for the derivative, like the

Gradient/Hessian [NA08] or derived from a certain edge model like the Canny or Canny-Deriche algorithms [Der87]. An example of edge detection, applied over an image can be seen in Figure 4.1(b).

Textures/Color. Texture/Color analysis searches to segment the image into zones that belong to the same object. One particularly common technique in human action recognition is the Skin detector. The most simple skin detector associates a volume in the HSV color space to skin, thus classifying pixel-by-pixel by checking if it is inside that volume [JR02]. Other techniques define probabilistic models to represent skin color [YA98]. See an example in Figure 4.1(c) of an skin detector using a volume in the HSV color space.

Disparity. Disparity is often the first stage for 3D reconstruction, as it can be computed using a set of images (at least two in stereo systems). The objective is to find for each pixel a scalar value, the disparity, to represent the distance from the cameras: the closer the object is, the higher the disparity appears. This problem can be understood as a variation of the computation of the optical flow, where the flow is computed over n images at the same time. Common algorithms to calculate disparity search for the correlation of several (shifted) images, a computation that can be performed locally examining the correlation only in a close neighborhood. Other approaches to compute disparity look for the correspondence of features like edges or corners in the set of images, for a review see [JJT91, SS02]. An example of disparity computed over a pair of images can be found in Figure 4.1(d), where we use a correlation based algorithm (SAD).

Optical flow. We discuss local approaches for the computation of the optical flow in Chapter 1, where more details can be found. We recall that the main idea in optical flow algorithms is to compute a displacement vector $\vec{v}(x, y, t)$ at each pixel location to, for example, minimize the difference between two consecutive images $I(x, y, t-1)$ and $I(x+u, y+v, t)$.

Global features

Global features provide information of the entire image or spatiotemporal volume. In human action recognition it is usual to see global approaches as human poses can be stereotyped, therefore templates can be generated. The main difference with local approaches is that the information of large neighborhoods or even the complete image/spatiotemporal volume is taken into account simultaneously. We present some representative global features extraction techniques:

Template matching. Some researchers consider that the human body can be simplified into a set of parts, like legs, arms, torso and head, where templates for each part can be built, see Figure 4.1(f). These templates can be hand crafted or automatically constructed from a training database to obtain invariant recognition (to scale/rotation) [HhYW05, RMR07, IF01]. Template matching usually provides likeliness functions for the position/scale/orientation of each body part in the complete image. These likeliness functions can be built using convolution, histograms and then the most likely position/scale/orientation can be found using RANSAC estimations (RANdom SAmple Consensus) or other statistical estimators, like MLE [Sap06] (Maximum-Likelihood Estimation).

Silhouettes. Silhouettes can be obtained from edge detection techniques [MTHC03], disparity maps [PF03] or combinations, see Figure 4.1(g). The objective is to obtain a connected area associated to the body of a person, thus it can be considered as template matching. However, silhouette detection can be found so often that we mention this approach separately. It is common to see this approach combined with other ones [WN99], like skin detectors

or template matching to obtain not only the silhouette of the person but also information about the position of each part of the body.

Local flow operators. Even though we have mentioned the optical flow as a local feature, some authors [Lap05, GP03] have considered that local patterns in the optical flow contain information to represent the complete spatiotemporal volume. Local flow patterns are small (smaller than the image) patches of the optical flow, that characterize a certain sequence of motion. The few authors that have proposed this kind of feature have either selected them by some statistical criteria (PCA, ICA) [JSWP07, Lap05] or taking inspiration for example in biology [GP03]. All these authors have proposed flow patterns like translation, rotation, contraction/expansion and discontinuities. Indeed, this idea has a bio-inspired support [GP03], where similar patterns have been reported in primates, see 4.2.1 for details. These features are not extensively used, and to our knowledge only few authors propose them in computer vision applications [GP03, JSWP07, Lap05]. An example of local optical flow pattern can be seen in Figure 4.1(h).

AdaBoost. The AdaBoost technique (Adaptative Boosting) is not a proper feature extraction technique but rather an algorithm to obtain a good classifier from a set of "weak" classifiers [VJ01]. It has been used in the context of classification, but also to combine several approaches to segment (classify) in computer vision. Particularly interesting is the derivation of local operators to detect faces, hands, and bodies from simple features [VJ01] and the detection of other body parts [MJB05]. We also notice the reported good results [LT11] of AdaBoost in realistic scenarios.

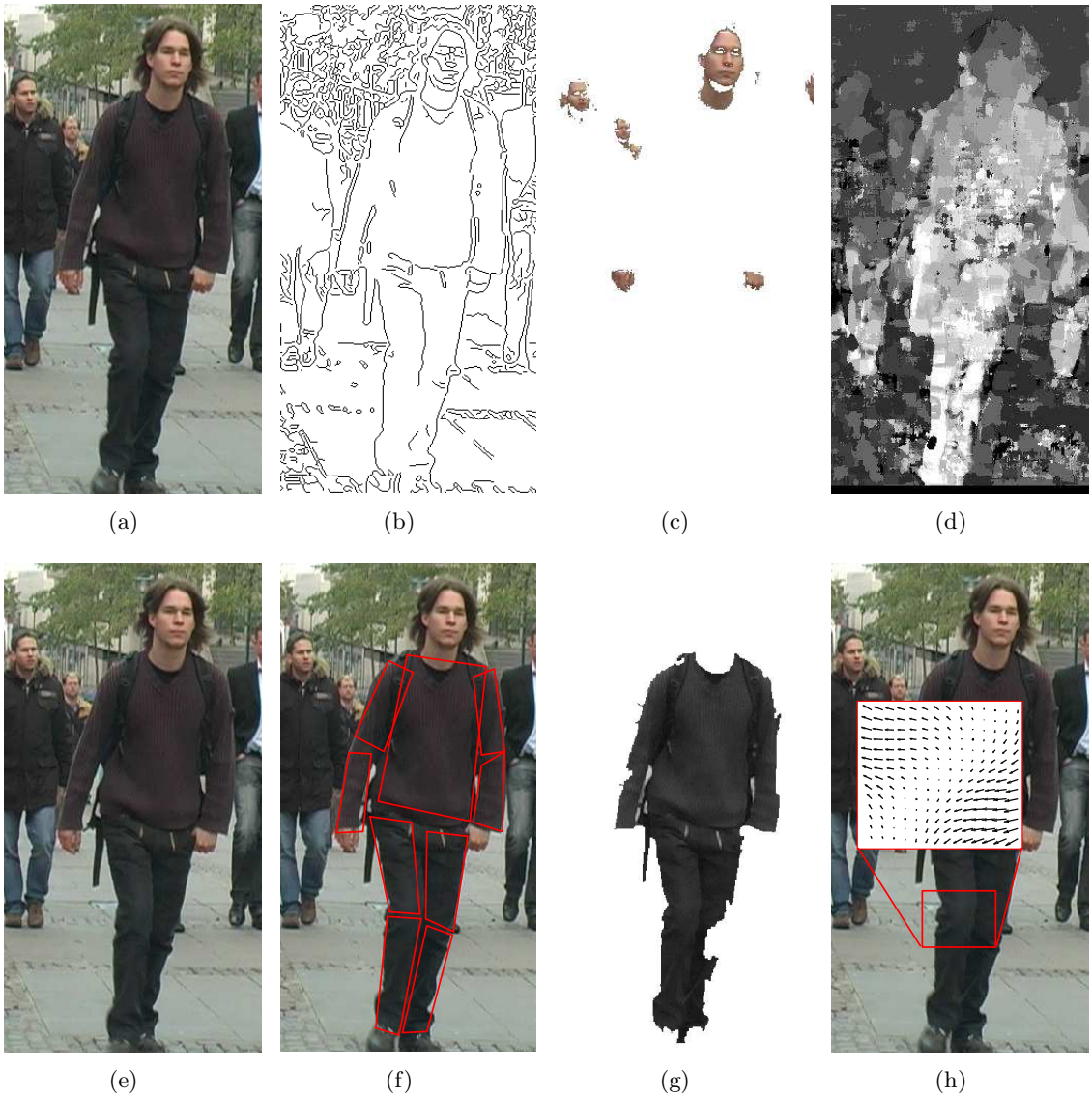
4.1.2 Sequence discrimination

Pattern retrieval, recognition, classification, discrimination and other not very informative terms refer to the same task: to label from a list of known objects an unknown object. Applied to videos of human motion, this idea becomes identifying different human actions. The human action recognition problem has been addressed using a wide variety of techniques [MG01, MHK06, Pop10] such as distances in some features space, HMM [5] states for each body pose or RBF [6] where each base represents a pose.

Human action recognition techniques are commonly separated into pose estimation and action recognition sub-problems [MG01, MHK06], where pose estimation looks for the current body configuration and action recognition deals with the interpretation of the set of poses. We also notice that action recognition can always be understood at different levels, since the same body action in different contexts can be identified as a different activity: e.g. a subject can bend down to pick up a flower or to exercise himself, in this case context information like the presence of flowers or the sportswear may help to make the difference, yet in both cases the subject bends down. To avoid this problem, we study only actions unrelated to the context and valid for any subject such as locomotive actions: to walk, to run, to jog and non-locomotive actions like: to clap, to fight, to wave.

5. Hidden Markov Models are statistical models of a Markov process, where the states of the systems are hidden but the output is observable. In this example, states may correspond to the current body pose and the visual features to observations.

6. Radial Basis function are a common function approximation with the form of a sum of radial basis functions. Applied to human motion, each base may correspond to a certain pose, thus any body configuration can be described with a large-enough number of body poses.



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

(e) (f) (g) (h)

FIGURE 4.1 – (a) and (e). Original images (left and right view of a stereo system). Local (b), (c), (d) and global features (f), (g), (h). (b) Canny contour detection, (c) HSV volume skin detector, (d) Disparity using SAD matching from (a) and (e). (f) Body parts detection. (g) Silhouette/Body detection using disparity and contours. (h) Local motion pattern projection, showing a high activation zone. The original image is from the TUD-Brussels public database [WWS09].

Pose estimation

Pose estimation searches for the body pose matching the current observation (an image). It can use local or global features, and the final result is generally the position/orientation of each joint of the body, more precisely the joints in the body model.

Model based estimation. Model based estimation uses anatomical information of the human body, typically the number of joints, their DoF (Degrees of Freedom) and the number of segments. These models can be composed from 8 joints in simple models [MTHC03], to over 30 in biomechanical applications [KRW90], like the one depicted in Figure 4.2(a). Once a body model is selected, a matching must be performed between the feature extraction and the model minimizing the distance of the prediction and the observation. To achieve this matching a wide variety of techniques have been proposed like: Bayesian methods [IF01], HMM approaches [KHM00] and distance based techniques [SH05].

Model-free estimation. Model-free estimation does not require an a priori anatomical model of the human body. Historically, model-free techniques are a more recent approach than model based ones [MHK06]. The representation can be a previously selected ensemble of body parts [MJB05] or an automatically learned one [CBK03], see an example in Figure 4.2(b). In this example a set of detectors (hands, foots, head) work simultaneously on the same image, and the set of detections represent implicitly the body pose. Hands, foots and head detectors can be built separately, and a certain structure [MJB05] (tree, Bayesian network) can be used to infer the body pose. The main advantage of model-free models over model-based ones is mainly the tolerance to occlusions [MHK06], because if some body part is not (or wrongly) detected, these models are still able to deliver likely body configurations.

Action recognition

Even if we suppose that pose estimation can be successfully performed, this is not sufficient to interpret the action and to be able to determine the current action. Action recognition requires, in many techniques, pose estimation information to infer the current action, but it is not the only approach. Other interesting techniques have been proposed in different research communities such as biomechanics, where it is well known that the information of only a few joints can characterize the movements of a subject. We do not study approaches based on scene interpretation, where the subject is supposed to be far enough to consider him as a point, as it has been also proposed in the literature.

Body-based recognition. Body based recognition uses the results of the pose estimation either with model-based or model-free results to label a given human action. The temporal set of pose configurations is used so as to classify either directly the action by using cross-correlation [RCKL06] against previous examples, or by first projecting into different spaces each pose information (using vector quantization techniques). Other authors work directly in the spatiotemporal volume (XYT), either by using slices [RB00], or histogram information [BD02]. For example in the case of the tHMI [BD02], each pixel that belongs to the silhouette in a sequence increases the frequency of the pixel in that motion. In figure 4.2(c), we see a periodic movement projected onto a feature space (in this particular case PCA, using 3 components), each point representing a pose: to identify an action in this space is to compare the distance between two of these curves.

Body parts recognition. Other than the mentioned holistic approaches, where the information of the complete body is used (even though body parts may be used to derive the

silhouette), a very different approach is to avoid pose estimation, and directly use some specific body part. In this approach we can find the use of body parts position, for example feet positions, to measure the time cycle for a leg or the time both feet are simultaneously in the ground [DT02]. We can find the inspiration of this technique from other research fields like biomechanics [Per92], where double/single stance (time when a foot touch the ground) time relative to gait cycle is standard in gait analysis. For example, in gait analysis, when the time of double stance goes close to zero the movement is faster, and at the limit it can be interpreted as running. In Figure 4.2(d) we can see a typical gait analysis for a walking person (the same as Figure 4.2(a)) using the distance to the floor of each foot. Another approach is to look for local descriptors in the spatiotemporal volume, to characterize a given action, and then use a metric [Lap05] to compare against a database of known movements.

4.1.3 Constraints and implementations

We have presented a general overview about how to estimate the pose and perform human action recognition, following current trends in computer vision. Despite their diversity, there are common aspects that we will discuss related to feature selection and to the way the presented approaches encode actions using these features.

With respect to feature selection, we notice that feature extraction mostly analyzes videos as sequences of images where a human pose must be estimated separately in each frame. Only few techniques use temporal information [Lap05, JSWP07]. Probably this is, at least, a poor choice as biomechanics clinical studies [KRW90] have shown that looking at a sparse set of joints trajectories while the subjects walk/run is sufficient to label movements or even to detect gait anomalies. Therefore, temporal information does provide useful information, sufficient to detect gait problems [KRW90, DOTG91] or even to identify persons [WTNH03]. Another aspect about feature extraction techniques is that many pattern recognition algorithms use multiple features at the same time, as a way to boost their performance, like in the AdaBoost technique [VJ01]. For example, when using silhouette information ambiguities can be reduced by using skin detectors or disparity maps [PF03, MTHC03] (from an stereo ring or a laser scanner). Other approaches integrate hands and face position [PF03, RMR07] with silhouette information to improve the pose estimation.

In the representation of sequences, much effort has been oriented towards action recognition using 3D information, i.e. using at least 2 cameras. Only recent efforts have been oriented towards monocular acquisition setups. The internal representations in these techniques tend to be 3D, to cope with rotation invariance and different view points [WN99, PF03, MTHC03, SH05]. This is surprising as even if recognition should be robust to rotations, all orientations do not need to be considered, i.e. it is not common to see people walking upside-down. A second interesting element in human action discrimination is the apparent disconnection between computer vision and clinical analysis techniques, where it is well known that only a few parameters allow the analysis of movements, for example, in gait analysis position and derivatives of knee position/orientation. Computer vision, with some rare exceptions [DT02], is much more centered on retrieving the complete body position, and then interpreting each pose along time. Timing seems also ignored: very few authors check [SvG08, RBBVdZ09] for example, how many frames are actually required to perform classification or how the classification evolves through time.

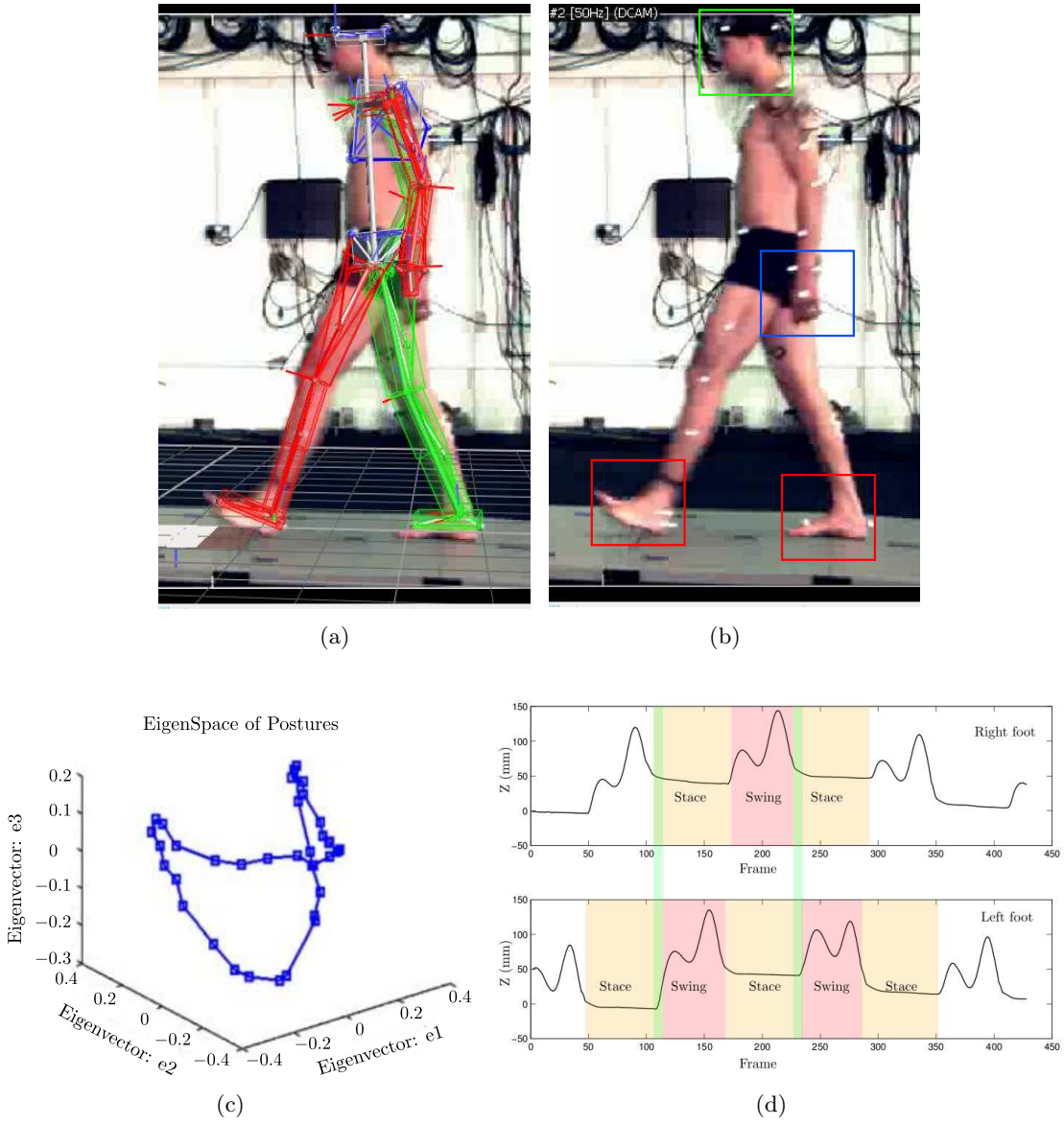


FIGURE 4.2 – (a) Pose estimation using an anatomical model and (b) without body model. (c) Action recognition based on full pose information and (d) looking at one body part behavior only (distance from the floor). Data used in (a), (b), (d) is from the example VICON© database. Data used in (c) is reproduced from [RI05]

4.2 Biology

Even though visual signals can be severely altered, action recognition is very robust in humans [Joh73, HK87, RFG01, GP03, PD07, BS07]. One of the first authors to precise this robustness was G. Johansson in the 70'. In his pioneer work, Johansson [Joh73] proposed the Point-Light stimuli where only the joints of an actor are marked. Then, he recorded different actions, such as to walk or to run and presented this video to other subjects. The striking result was that only marking the joints of the actor (or with even random PL as shown recently by Casile et al [BP94]) is sufficient to allow good pattern classification in humans. Hence, very sparse local information already allows good action recognition.

Numerous experiments related to the perception of human and animal motion are available, starting with the photographic work of Muybridge [Muy55] at the end of the XIXth century and later by Johansson, yet there is still a vivid debate interpreting these results [Joh73, GP03, BS07]. From the perceptive experiences proposed by Johansson himself, most of these experiences come from experimental psychology as the classification task is associated to higher areas of the brain where it is difficult to target a specific neuron or area. More recently, studies have used medical imaging techniques like fMRI to identify different zones of activation/inactivation while the subject sees human actions. Despite these efforts, the exact features that allow classification as well as the precise mechanism for coding and retrieval remain in discussion [BS07].

4.2.1 Extracting Features: the visual cortex

As we have seen in Chapter 1, the visual cortex can be described as two streams of information: the ventral pathway carrying static features information and the dorsal pathway processing movement and spatial location [CW03, KSJ00]. This view is currently debated as many interactions exist between the flows. The separation appears fuzzier as more studies become available. However, to organize our overview about the main related areas in human action recognition we still keep this usual division. We will start with a general overview of the two streams where we provide a more detailed description than in Chapter 1. Then in 4.2.2 we focus specifically onto the human action recognition task in primates.

Ventral pathway

Receiving most of its input signal from the P pathway (color, high resolution information) coming from the retina through the LGN, the ventral pathway is associated to static information thus being called the "what pathway". Most information first goes into area V1 connecting to layer 4, see Figure 4.3. Area V1 receives information from several paths coming from the thalamus (P,M pathways) at different layers, such as orientation/colors (layers 2,3,4A), disparity and motion (layer 4A). Area V2 receives its main input from layer 3A in area V1, where processing of form and color has been reported, with a binocular component and most of its orientation selective neurons being end-stopped [7] [HL87, LKM94]. The receptive fields in V2 are larger than in V1 but many of the receptive field properties found in V1 can be found in V2. Other than the receptive field size, V2 performs interactions within orientation and the near surrounds [LKM94], for example some illusory contour responses have been reported in V2 but not in V1 [PvdH89, HVE00]. Area V2 projects mainly into area V4 (and MT), in average with a receptive field 4 to 7 times bigger than in V1 [PPRF02]. In V4, medium complexity figures have been reported as receptive fields (like concave/convex contours and corners), with some position invariance [PC01].

7. End-stopped neurons are sensitive to motion but only of corners, thus not attained by the aperture problem.

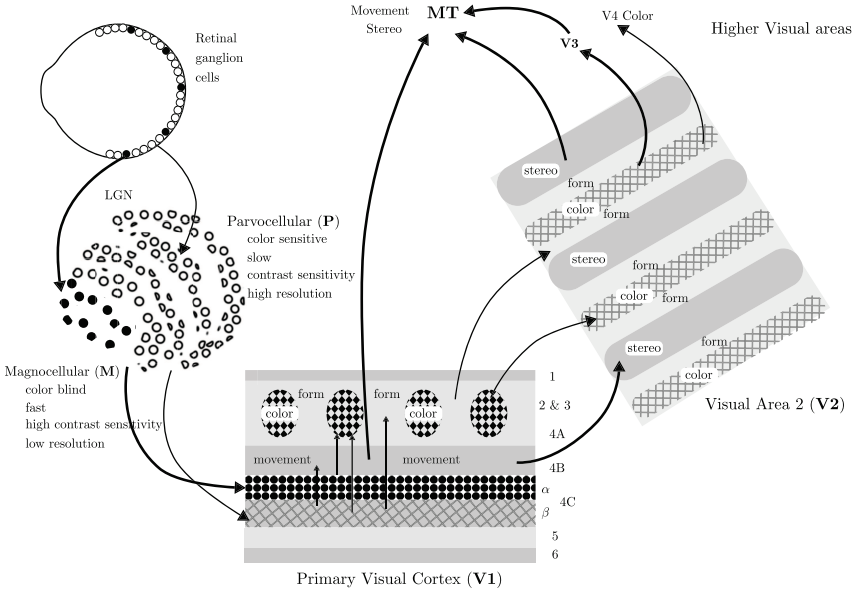


FIGURE 4.3 – Layers in the visual cortex. Thick arrows represent the dorsal pathway flow. Layer I is closer to the cortical surface (reproduced from [CW03]).

The IT area (Infero Temporal) is the next stage of processing in the ventral pathway, with reports [Rol00] of complex static features with receptive fields such as: hands, faces and bodies. IT receives its main input from area V4. IT has neurons with receptive fields larger than in V4 by a factor of 4 [Rol00]. Studying a series of gradually less realistic features, for example the picture of a hand, then a cartoon hand and a few geometric elements of the hand, area IT shows progressively less activation [PD07]. It is not clear, however, how neurons in this area achieve this selectivity. A first model proposes a hierarchical processing, where IT takes geometric "primitives" from V4 and builds the complex detectors, but this is not the only model, for a review see [RP00]. More recent studies in face recognition report that there are also neurons sensitive at the same time to two or three visual primitives [FTL09], for example simultaneously coding eyes and mouth characteristics.

Receiving input from the IT region and in the context of biological motion areas, STS (Superior Temporal Sulcus) and EBA (Extrastriate Body Area) are supposed to receive inputs from pre-motor areas like F5, or at least to present correlated activation. The receptive field size of area STS is even larger than V4/IT covering almost the complete receptive field. STS area is activated when articulated movements (like human motion) are presented. Receptive fields in area IT have been compared to a "pose" detector where neurons are activated by poses while observing human actions [GP03], and areas STS/F5 are supposed to be activated by temporal sequences of these activations. Area STS is also associated to dorsal processing, so a convergence mechanism may link both streams [PVVO05].

Dorsal pathway

Areas V1, V2, V3, MT, MST are the primary regions of the dorsal pathway receiving input mainly from the M pathway (achromatic information), where information related to movement and spatial location has been reported by single and multiple neuron activation protocols [KSJ00, CW03]. Areas V1, V2, V3 and MT are retinotopically organized with reported feed-forward and feedback connections in the dorsal sense among them [HPL$^+$98, HJG$^+$01]. V1 and V2 areas are shared between the ventral and dorsal pathways. It should be noticed that in each region there are layers, and thus information is still segregated within this network, see Figure 4.3. The difference (in terms of regions) between the dorsal and ventral pathways starts in V3, a region found bi-hemispherically in the occipital lobe composed by neurons with a receptive field larger than in V2. One of the distinctive characteristics of V3 is its chromatic processing and integration of motion, integrating information from the M and P channels coming from the thalamus [GKL97] and V1. The V3 processing illustrates how the segregation between dorsal and ventral is not evident, as interactions exist in almost each region. Region MT/V5 in monkeys (hMT/V5 in humans) is specialized in the detection of motion with "strong" inputs from regions V1 and V3, exhibiting tuning for orientation and speed, and activation with illusory contours, chromatic movement, and solving the aperture problem [BB05]. Depending on the reference, the receptive field is 4 to 10 times bigger than in V1 [SH98].

After motion is computed in this specialized network (V1, V2, V3, MT/V5) [8], the information is processed in the MST (Medial Superior Temporal) region considered a satellite area in monkeys [KSJ00], and in the hMT/V5+ complex in humans which includes the equivalent to the MST area. These areas are located in the boundaries of the visual cortex, see Figure 4.2.1. The topology of MST is retinotopic, but with large receptive fields from 7 to 70 degrees [DW95], with units covering up to a quadrant of the visual field. About their receptive field, experimental evidence indicates that it corresponds to local patterns of movement, like translations, rotations, contractions, spirals [THS$^+$86] but its topology is less clear than for example, in areas V1/MT, where a clear orientation topology (pin-hole organization) has been reported. Also, strong influences have been reported from eyes movements [DW95], and the area MST is commonly associated to ego-motion extraction because of its prominent vestibular influences [Duf98].

In the 50's Gibson [Gib50] enunciates that the optical flow can be always decomposed into a set of: rotations, translations and divergence/convergence patterns, and later Koenderink and colleagues [Koe86] theoretically prove this idea for small planar patches. This theoretical result had a great influence on biologists; for them it was reasonable to search for such kind of receptive fields in areas such as MST, yet experimental results have shown a somehow more complicated scenario: MST receptive fields are shifted by eye movements [DW95], and it is also known that vestibular areas (encoding information about the position of the head/body) have a large influence on the MST area [Duf98]. Thus, MST seems to be an area of multiple interactions and interpreting its functionality as passive feature extraction is probably wrong. However, clinical lesions in monkeys revealed that damages of the MST region degrade the detection of motion patterns especially in the presence of noise [PM94], by consequence the area seems strongly related to feature extraction in the biological motion task.

It is not clear, from the literature, if the equivalent to the IT area (ventral pathway), where neurons activation has been reported for static body shape stimuli, exists for the dorsal pathway.

8. see our presented model in Chapter 2

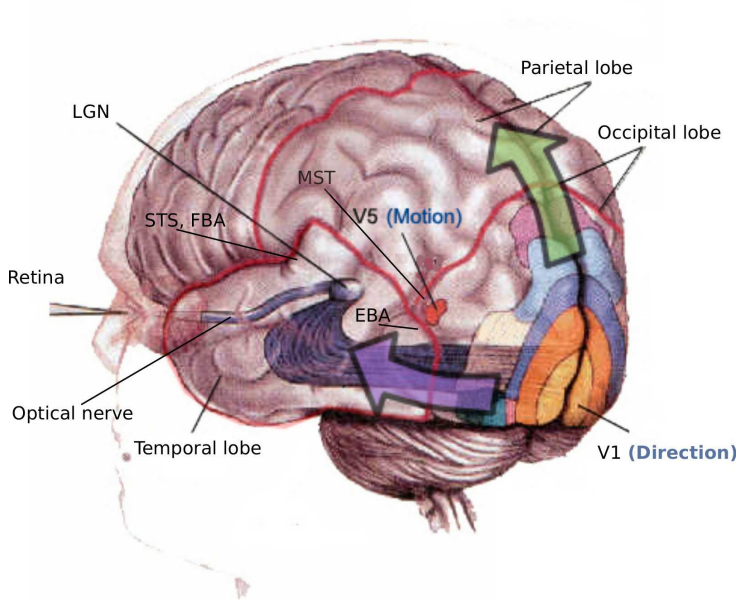


FIGURE 4.4 – A schematic 3D representation of the visual cortex and area locations. Green and purple arrows represent the dorsal and ventral pathway flow respectively.

Some authors hypothesize that the equivalent area may exist [GP03] and others that, instead, a strong interaction between IT and hMT/V5+ (in humans) could exist [PVVO05].

4.2.2 Discriminating patterns: related areas

The selection and interaction of features to encode visual stimuli is not well understood, but the interactions between the ventral and dorsal pathways seem systematic rather than isolated. More precisely, there is a debate about how temporal sequences (as human movement) are encoded in this context, which features are important to discriminate sequences and the relative influence of the ventral and dorsal pathways in this task [PVVO05, PD07]. However, a specialized brain mechanism dedicated to the recognition of biological movements (humans, animals) has been accepted in the last decades, as patients with parietal lesions can discriminate motion but not PL stimuli [AL00], for example. The precise nature of this mechanism remains, however, unknown.

Evidence exists [PD07] that areas STS (Superior Temporal Sulcus), FBA (Fusiform Body Area), F5 (Premotor area) and EBA (Extrastriate body area), see Figure 4.2.1, are sensitive to human actions such as to walk, but with some important differences. Area EBA is sensitive to static body configurations and activated by pre-motor activity (F5) [BS07]. Area FBA is located in the fusiform gyrus (anatomically far from EBA/STS/F5) and overlapping with area FFA (Fusiform Face area), commonly associated to face identification [KY06], thus it may be associated to the identification of subjects through body movement. The STS region presents a higher activation when both static and movement cues are available (and coherent).

According to the above discussion, there is evidence about areas sensitive to static features or "snapshots" such as body postures, face expressions and hand gestures in the ventral pathway, mainly IT (static features), and in areas such as EBA. These areas have been identified mainly

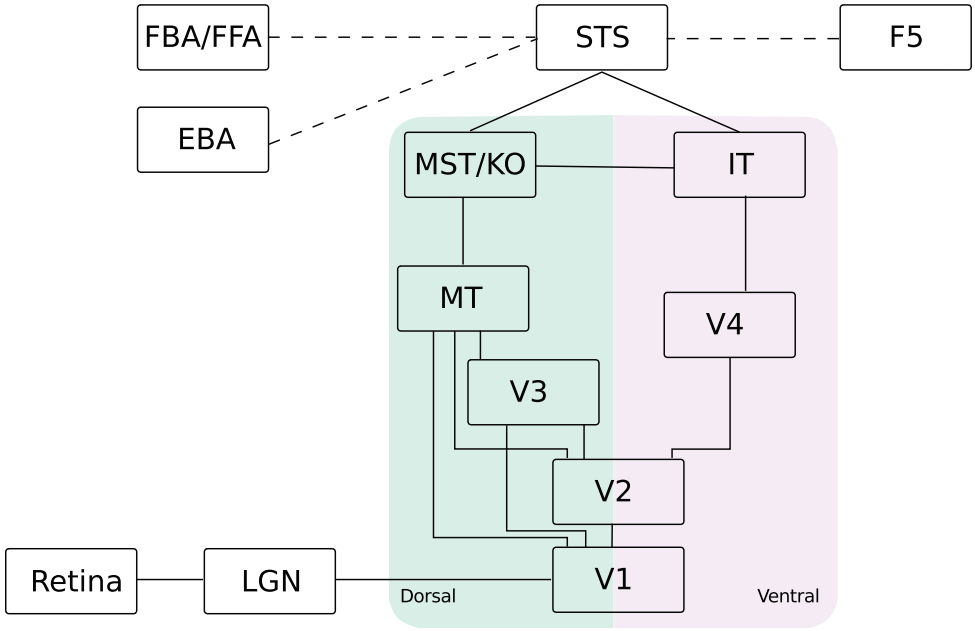


FIGURE 4.5 – Connectivity diagram of the main cortical areas related to the processing of biological motion. Dotted lines stand for correlated activity.

by fMRI/PET [9]/TMS [10] studies, so general information is available but recordings of single units are difficult to find, specially in humans. The activation of neurons sensitive to human actions by motion information only (dorsal pathway) is not clear. However it seems likely as the PL stimuli and its many variations have shown. Furthermore, in [SP09] the same brain areas have shown an activation with static and dynamic face expressions, greater for the latter case. All this could predict the existence of dorsal (movement) only snapshot areas or units [PD07]. Following the hypothesis that the motion is sufficient to perform classification whether it is integrated or not with static information [GP03], a possible mechanism has been proposed to perform the coding: single 2D snapshot units activated with the motion covering the complete visual field [PD07]. We call this hypothesis the explicit snapshots theory, and its major prediction is the existence of snapshot units (neurons) activated only by movement information, where each unit should be activated by a certain pose.

In our description about the ventral and dorsal processing of information, we have implicitly said that human recognition shows a uniform performance at any point in the visual field, but this is not the case. In fact, peripheral areas of the visual field are significantly less sensitive to human actions as experimental evidence indicates [BS07]. We are not very aware of this restriction, and it can be interpreted as a mechanism of further simplification, where complex features are only processed in foveal areas. We notice, however, that the eye processes are not static but dynamical (with saccadic movements up to 1000 degrees/second [FB83]), thus foveal processing should not

9. Positron Emission Tomography, a nuclear medicine imaging technique able to produce 3D images of some process in the brain.

10. Transcranial Magnetic Stimulation induces depolarization in neurons, by rapidly changing a magnetic field, and thus measuring this current to produce 3D images.

be considered as a strong limitation for the recognition of humans actions.

So far, we have described the general architecture of what is believed to be the recognition of human actions, describing involved areas and whenever possible neurons responses. However, there is also evidence from experimental psychology and computation time describing action recognition, to characterize in general terms the recognition of biological motion that we will now present.

Temporal sensitivity. It has been shown that action recognition is extremely sensitive to temporal variations [BS07, GP03]. Taking as an example the PL stimuli, in a cyclic movement (e.g. walking), if the spatial structure is preserved but each joint is shifted in phase with respect to the gait-cycle, the discrimination is significantly diminished [BP94]. With respect to the global temporal sensitivity, time-compressing a sequence seems to increase the sensitivity in the subject identification task using the PL stimuli [HP00]. This latter work shows that even if there is a "normal" speed, higher speeds or exaggerations are not harder to recognize, and even more, they seem easier to process. It is also well known that only a few "frames" are sufficient to identify human motion, where the work of Reid et al [RBBVdZ09] pushes this idea to its limit, to what he calls "implicit" sequences: the movement implied by just one frame.

View dependency. Experimental psychology has also explored the orientation invariance of action recognition, concluding that recognition of visual patterns depends on the angle of view of the observer. Experimental evidence shows that the same subject decreases performance when the presented pattern is rotated, improving his performance through experience. The usual tolerance is about 20° [PP03] before the recognition rate drops. This performance cannot be clearly associated to neither the ventral nor the dorsal path as the protocols we refer to involve subjects and not specific zones nor single neurons. These experiences unveil the question of whether the internal representation of motion is 3D. Experiments by [BBS98] indicate that at least for the PL stimuli (mostly dorsal information), a 2D representation is sufficient to explain brain coding schemes for body actions. In their experiences, Buelthoff and colleagues show that distortion in the direction of sight of a walking pattern (in depth) does not affect recognition, concluding that 2D rather than a 3D representation of motion seems more likely. This remains in discussion, because 3D representations could still exist with intermediate 2D projections (top-down). To summarize about view dependency, these two experiments, view dependency proposed by [PP03] and the scrambled PL stimuli proposed by [BBS98], strongly suggest that a 2D coding scheme is more likely.

Response latency. Another aspect related to the computation of feature extraction characterizes the recognition of visual patterns: time. As we have studied in the first part of this thesis, motion estimation can be understood as a distributed computation, but this approach requires additional time to eliminate ambiguities, for example to solve the aperture problem (around 75 ms after the first response in the human brain [SMM10], and from 4 iterations in our model), see Figure 4.6 for reference values of the time to first response in different brain areas involved in the recognition of complex motion patterns. At the same time, there is an intrinsic error from the discrete nature of the speed sampling, that is relative to the estimated speed (around 5%), thus at higher speed there is a higher absolute error. Despite this, tasks such as biological motion recognition (to walk, to run, to jump) where local motion information is a crucial input, take very short time, providing for example categorization starting from 120 ms to 200 ms [PD07] depending on the reference. Also, we notice that some studies, like [SLC97], show that biological movement recognition

can be performed even if the aperture problem is present, thus this kind of inaccuracy in the detection does not seem crucial for the biological motion recognition. Then, this kind of cognitive task cannot entirely rely on the precise feature extraction, but on aspects that can be computed almost directly and that are tolerant to errors that we know are intrinsic to the local estimation of movement, such as the aperture and speed estimation with a certain relative error.

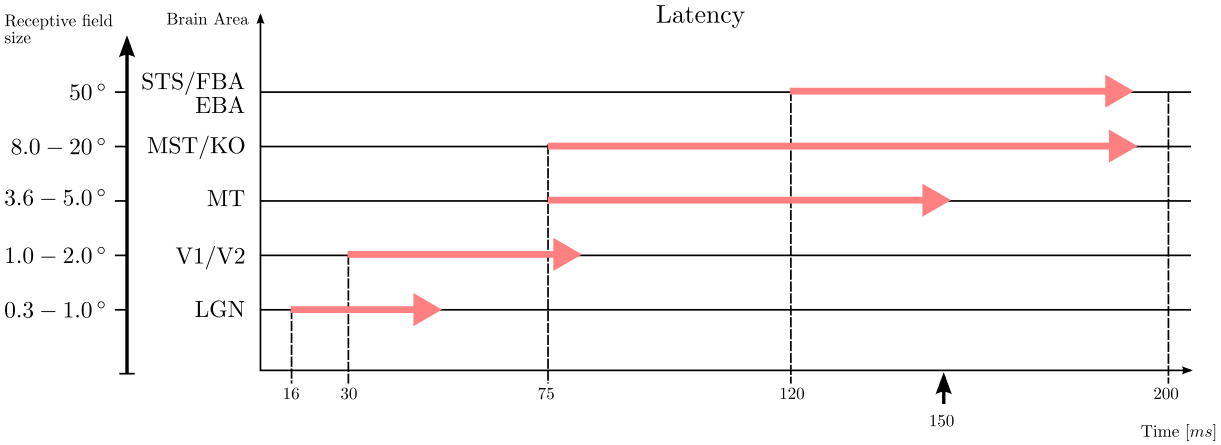


FIGURE 4.6 – Latency of different brain areas involved in the recognition of visual patterns (biological motion). The latency is the time for the first activation of the area after the visual stimuli is presented. The activity is not instantaneous. It can start and last several hundred ms after the initial stimuli in some brain areas. Note that the size of the red arrows shows that a certain brain area presents activity, at least in that period (could be longer). The small black arrow indicates the time when MT is believed to deliver a disambiguated response. The left column gives the receptive field of each area (see references). The construction of the diagram (latency, duration and receptive field) was based on the following references: LGN [MGA+99], V1 [MG92], MT [SMM05, SMM10], MST [FL97], STS [PD07]

4.2.3 Feature selection and representation

Different studies [TG08, CG05] indicate that the most relevant feature to perform classification of human movement sequences (at least for the PL stimuli) is the local motion (processed in areas V1/V3/MT/MST, see Figure 4.4). This was tested using variations of the PL stimuli where an inter-stimuli blank interval was used between frames to show that recognition is severely diminished if local features are disrupted [MRS92]. However, this argument has been debated by [BP94] in several experiments where the local structure was disrupted by marking points along each bone rather than the joints, yet the discrimination performance did not drop as expected (lower, but not significantly lower). More recently, a similar degree of discrimination has been achieved with static PL stimuli figures implying movement [RBBVdZ09]. These results show that local motion is a key feature, but top-down (and probably form) influences also play an important role. The interactions between position (dorsal path information: movement, position) and static information (ventral path information: geometry, colors) are not clear. As an example of this interaction, fMRI studies have shown that the same brain areas are sensitive to what is classically called the ventral path information, and to the dorsal one [SP09]. Interestingly,

this last input produces greater activations, at least for facial gestures [SP09]. Local movement information seems then to be a relevant feature in order to classify visual sequence, but not the only possible mechanism in place, as suggested by the likely existence of important top-down and form influences. Also, the precise movement information seems highly unlikely to be required to identify complex visual sequences, thus the spatial structure of the movement is more relevant, in other words: where there is movement along time.

The common view of two separated processing flows (ventral and dorsal), has been discussed since the 90' and more recently in the context of visual sequences classification for [SP09] faces. In their work Schultz et al. show that common areas are activated for static and moving faces, and even that the moving stimuli generate a larger activation of areas such as STS, FFA, OFA (Occipital Face Area). Certainly, this evidence holds for faces, but similar mechanisms for bodies may be hypothesized. The fMRI experiences by [SP09] confirm that there are not two separated mechanisms to match temporal sequences in STS, as also proposed before by [GP03]. Yet, even with a temporal matching that relays in a shared mechanism between the dorsal and ventral streams, the question of which local features are used to encode sequences of biological motion in the dorsal stream remains open. It is unclear from biology if the sequences are effectively encoded in explicit snapshots units and thus by global features, as the main theory currently predicts [GP03], or by interactions among local features. The interaction between local features by means of an implicit movement model (body model), could be a possible brain mechanism to encode complex visual sequences. If this kind of model is biologically plausible, some of the properties for the coding of visual sequences that we have studied like view variance, rapid classification, tolerance to time-warping should be at least retrieved by computer simulations.

4.3 Computer vision and biology

We have presented various techniques in computer vision and experimental evidence from biology to describe the kind of features used in recognizing human actions. Several questions recur under both perspectives: the feature selection to perform the recognition and the most appropriate architecture (2D or 3D) to represent biological motion patterns.

Computer vision techniques neglect temporal features and only a few studies [Lap05, JSWP07] take into account temporal features for the representation of sequences. It is well established in biology that very diminished signals, such as PL stimuli, still allow human action recognition. The study of these signals has inspired movement analysis in biomechanics, such as gait analysis. This suggests that local temporal features (such as local patterns of the optical flow) could be very important to recognize human motions. The same ideas can be found in several effective approaches of computer vision that are based on local features (sometimes combined with techniques like AdaBoost). These techniques have been very successful in detecting heads, hands and bodies by selecting a set of local features that together perform the detection. This kind of statistical criteria may be useful to predict shapes for the receptive fields in areas like MST or STS, where complete cortical mappings are not available. They could also be useful to identify the precise features that allow human recognition, considering that statistical criteria may be related to the receptive field in areas MST/STS. Independently of the features, experiments from biology, where temporal local information is available, report superior recognition performance. This is particularly interesting in computer vision, where feature selection is a problem in itself.

The best architecture and the precise coding mechanism for visual discrimination are not clear in either computer vision or in biology. Two radically different approaches can be found in computer vision: 2D or 3D encoding and the choice is not easy. This presents an interesting

question to biology: do humans encode visual actions in 2D, 3D or is there a different mechanism? And at the same time, what are the properties of such a system? The link among the encoding strategy (2D/3D), the approach (statistical/probabilistic), and the properties of the recognition is not obvious. The brain may provide an interesting inspiration and link between both ideas: the encoding strategy and the characteristic of the recognition. One of the possible advantages of a bio-inspired architecture for the recognition of human motion is that feature extraction and even classification could be performed in a distributed architecture. This could potentially inspire new, good-performance algorithms for computer vision to encode and recognize spatiotemporal sequences.

Chapter 5 presents a model of how a set of local features can provide a distributed mechanism for temporal sequence discrimination. This model supports the idea that 2D local movement representation is sufficient to discriminate visual sequences. We propose the spatial structure of movement over time rather than the precise movement as a possible mechanism to explain how the brain encodes complex visual patterns. Later in Chapter 6, we present experiments performed with our discrimination model using real video sequences to verify the effectiveness of our approach. In the same chapter, we also evaluate the discrimination of our model using two features: based on raw optical flow or local optical flow patterns (translations, discontinuities, rotations) associated with MST area. Feature comparison allows understanding the effects of choosing between these two kinds of features in the discrimination task.

5 Temporal pattern discrimination

Contents

Humans and animals interpret complex visual stimuli such as body movements and face gestures almost without any effort. Among other tasks, the discrimination of visual sequences without context is a key problem to understand both encoding and retrieving of spatiotemporal patterns in the human brain. Historically, this problem has interested biologists but recently there is an increasing interest from the computer science community to model and take inspirations from these ideas. In this chapter we present a model for the discrimination of visual sequences, based on the Continuum Neural Field Theory. Our model takes into account several properties exhibited by experimental psychophysics and physiology. The presented model shows how sparse spatial encoding of spatiotemporal sequences could be sufficient to explain some of these properties, such as classification with partial sequence information in space or time, and show that local motion information with a very simple body representation is sufficient to perform sequence discrimination. We can also encode a temporal sequence with a single population of units, without the need for explicit "snapshots" or estimations at each time instant. The model we introduce in this chapter shows from both a theoretical and an experimental point of view how a population of units may encode a global pattern of activation by a distributed mechanism.

5.1 Temporal pattern encoding

In the precedent chapter, we described experiments about different aspects of the encoding of biological motion. In this section, we give an overview of the main plausible biological computational models for temporal pattern discrimination and discuss them in relation to the experiments described in the previous chapter. This review will help to introduce our model in section 5.2.

5.1.1 Local features

Some researchers [MRS92, JSWP07] have argued that local features characterize visual sequences. Mather et al propose that local features completely describe biological motion, in particular extremity movements (like hands, foots). The main prediction of the suggested mechanism (there is no practical implementation) is that local features are critical for biological motion, an idea that has been debated by the work of Thorton et al [TI98], where no significant performance reduction is observed, when local features are perturbed (by means of ISI). Even considering this, we notice the works of [Lap05, JSWP07, EMVK09] that have shown from a computational perspective how the combination of local motion features with a classifier (like SVM or other metrics) can provide an effective way to encode visual sequences.

5.1.2 Local structures

Historically, the first model was proposed by Johansson [Joh73] and modified later by [WBN85], where it was suggested that the structure, in particular the position of joint pairs (bones), can encode a biological motion sequence in a model called "Visual Vector Analysis". To our knowledge, there is no computational implementation of this model, but we mention it because it was a first proposed mechanism. This hierarchical view of biological sequences was greatly inspired by the Point-Light stimuli (PL), and it has support from experiences where the relation among joints was shown to be important. However, this model predicts that joints detection is critical. Indeed, one of the PL perturbations proposed by [BP94] shows that if middle bone positions (instead of the joints) are marked, the recognition rate drops, but not significantly. As a consequence, the relation among joints does not seem critical to perform recognition, arguing against the visual vector analysis model.

5.1.3 Snapshots

The work of Giese et al [GP03] presents a complete computational model of the recognition of biological motion. It considers the dorsal and ventral pathways, and it argues that a parallel recognition can be performed with the information of either one of the two pathways information (form or motion). Considering the dorsal information, this model proposes a feed forward processing where local motion information (V1/MT) is further processed to extract discontinuities and translations patterns (MST). The main prediction of this model is the existence of neurons than represent a visual "snapshot" based only on dorsal information, thus describing a temporal sequence as a series of snapshots of the complete visual field. After the spatial matching, or pose estimation, a sequence matching mechanism based on the neural field theory is proposed by the model of Giese and colleagues. This model can reproduce a number of experiments, to name a few: view dependency, rapid discrimination, recognition using the PL stimuli, among others. The main discussion of this model is the theory of "snapshots" neurons, because this kind of neurons has not been observed in the brain (it is a model prediction). Additionally, this model cannot

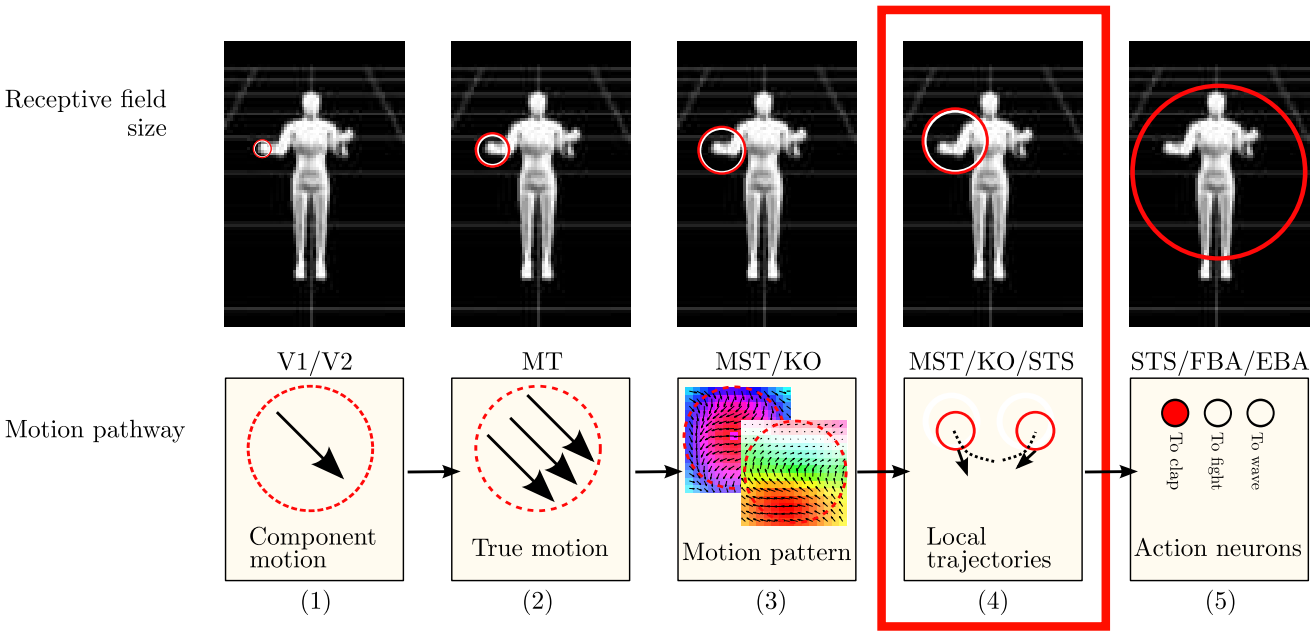


FIGURE 5.1 – Dorsal areas and our view of the processing. Our main modeling effort, developed in this chapter, is in stage 4.

explain experiments such as [TI98], where the local structure is perturbed but the performance does not drop significantly, as a feed-forward model predicts.

As we have described, at least three different approaches exist to explain the coding of temporal patterns for biological motion, in all of them local features have proved to be important. Complete visual field neurons, the main prediction of the snapshots model, have not been observed in the brain, thus other hypotheses may be studied. In this thesis, I argue that a set of local features, organized to describe a temporal sequence, can also explain many of the observed properties of biological motion recognition in the brain. To build this model, we consider an extension of the model proposed by Giese et al [GP03] into 2D to describe sequences as sets of local features, see stage 4 in Figure 5.1.

Our model highlights that local features, combined with a simple body representation to predict local features, can provide an effective visual discrimination mechanism, able to reproduce several properties observed in the recognition of biological motion. However, we cannot discard the snapshot model; our model proposes an alternative hypothesis, that to our knowledge can be supported by the wide variety of receptive field size of neurons in areas such as MST, providing enough support for our proposed distributed (implicit) representation. The model we propose could explain why pose neurons have not been observed based on dorsal information: a distributed representation of complex visual patterns will be much more difficult to observe.

5.2 Model description

Assuming that a temporal visual sequence can be reduced into sets of trajectories in time, following the above discussion, we consider that if we are able to encode these trajectories, we can encode the complete sequences. This implies that we suppose that we are able to know the

positions of these points in time, and we want to differentiate trajectories in time. For example in biological motion, these trajectories can be associated to joints trajectories. In this chapter, first we assume in our modeling that it is possible to do the local feature extraction to obtain several spatiotemporal trajectories, idea that we evaluate in the next chapter, and we focus on the pattern (sequence) classification, to encode the spatiotemporal structure of the complex pattern. In order to differentiate spatial trajectories we extend a computational model based on the Continuum Neural Field Theory (CNFT) [Ama77, Tay99, XG02], where the visual input is mapped onto a 2D population of units or neurons. We start with a description of the theory, a mathematical framework to model populations of neurons. Then we introduce a modification into the CNFT model so as to represent spatiotemporal trajectories, studying the dynamics of the model and parameter adjustment theoretically, to verify the coding properties at the end of the chapter.

5.2.1 Continuum Neural Field Theory (CNFT)

Given the complexity of the brain, it is important to properly choose a scale of observation. We can observe the brain as series of biochemical reactions at the level of neurons and synapses, or single neurons activity (spikes) or even the average activity of large ensembles of neurons. As we are interested in a cognitive task where we observe activity in different brain areas, we use a theory capable to handle this last scale of observation, the CNFT.

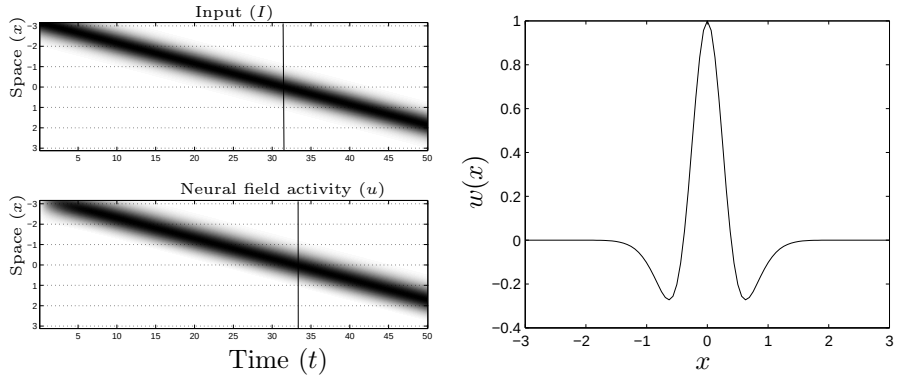
In the CNFT, the activity of neuron populations (membrane potential) is approximated by their mean firing rates, where populations are continuous neural sheets, with no transmission delay among units [Ama77] and a linear dependency with respect to the input strength. With these considerations, the membrane potential m follows Eq. 5.1:

$$\frac{\partial u(x,t)}{\partial t} + \tau u(x,t) = \int_{\Omega} w(|x' - x|) f[u(x',t)] dx' + E(x,t) + h \tag{5.1}$$

where E is the external input activity (retinal input in the case of the visual cortex), h is the neuron threshold, f is the neuron activation function and the integration is over the full set of neurons Ω. Of course, the cortex looks more like a 2D manifold rather than 1D, but for the sake of simplicity we start with the 1D case. One very studied type of solution for u are different forms of "bumps" of activity. These bumps can be used to model patterns of self-sustained activity, see Figure 5.2(a), propagation of activity fronts [Coo05] or competition among populations [Tay99], these various patterns having been observed in the human cortex and other brain structures, as described in the review of Wu et al [WXC08].

As we will focus later on the dynamics of activity fronts (or input-lock activity), we start by describing this kind of dynamics in the neural field. Activity fronts have been observed before in cortical preparations and they have been studied from a theoretical point of view [Coo05]. The front behavior for the neural field is dominated by the external input activity $E(x,t)$ that drives the activity of the neuron populations $u(x,t)$, moving at the same speed as the input, but with a certain delay, see Figure 5.2(a). To observe this behavior, several shapes for $w(|x' - x|)$ can be used, for example a symmetric Mexican-hat connectivity or Gaussian difference can be used, to have local excitation and long-range inhibition, see Figure 5.2(b).

The dynamics of the system can be illustrated as in Figure 5.3, where an input moving at constant speed drives the activity in u. We notice that if we measure the total activity, that we note $A(t)$ (over time and space), an input moving at v will be indistinguishable from another moving at $-v$, because the connectivity is symmetric. The same traveling-wave behavior can be observed in 2D, see Figure 5.5, considering the Mexican-hat kernel in 2D (Figure 5.4). In this



(a) (b)

FIGURE 5.2 – (a) Moving input (upper panel) and neural activity u (lower panel) that "follows" the input activity using the kernel defined in (b). The delay time of the neural field is given by the vertical line in each panel. The input is a normalized Gaussian function and the kernel (w) is a Mexican-hat. (b) Difference of Gaussians kernel function, or Mexican-hat used in (a) as kernel function.

last example, two simultaneous and opposite traveling waves move around a circle. As the kernel is symmetric the amplitude of both bumps is identical.

The propagation of activity fronts have been proposed, for example, as a robust distributed mechanism to perform video tracking. In this case, as the system reacts in the same way for inputs moving at different speeds, due to the symmetry of the system, it can be applied to video targets moving in different directions and speeds, where the front represents the target to be followed. Also, neural fields have been used to formally study the dynamical properties of neural populations in the brain, like the stability of the system, the number of bumps the populations may have or even other patterns that can exist in this kind of model.

5.2.2 Asymmetric neural field (ACNFT)

In 5.2.1 we have explained the main ideas of the CNFT theory, because we plan to use this framework to model the sequence discrimination task. We consider now a simple 1D sequence, where the only relevant property to encode is speed. We first transform Eq. 5.1 to remove the non-linearity from the integral, following the idea by Xie et al [XG02], see Appendix A.1.1 for more details. Also, to avoid border effects we consider the neural field in a ring, thus integrating over a circular domain,

$$\frac{\partial m(\theta,t)}{\partial t} + \tau m(\theta,t) = f\left[\int_{-\pi}^{\pi} w(\theta'-\theta)m(\theta',t)d\theta' + I(\theta,t)\right] \tag{5.2}$$

In Eq. 5.2, we have introduced a slightly modified version of the neural field model, that can also exhibit input-lock traveling [HS98] solutions as the original version. The main difference with the classical neural field equation in Eq. 5.1, is the relation between the variation of the neural activity and the external input: in Eq. 5.2 there is no variation at saturation levels. We now address the question of how to bias the traveling wave to have a bigger amplitude for a certain input speed to encode $I(\theta,t)$ moving at a certain speed.

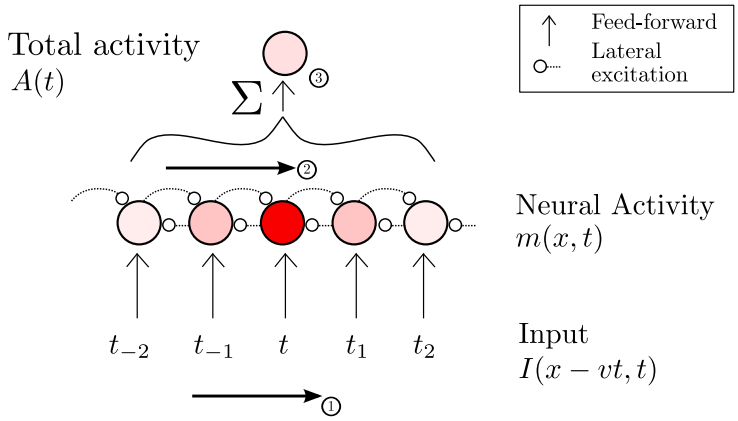


FIGURE 5.3 – Symmetric neural field dynamics. Given an input activity moving at speed v (1), the neural field can be tuned (shape of $w(x'-x)$) to have a wave behavior moving at the speed of the input (2). In this case the total activity (3) will be the same from an input, for example, moving at $-v$. In the Mexican-hat form for $w(x'-x)$ there are also inhibitory long-range lateral connections (not shown) that are also symmetrical.

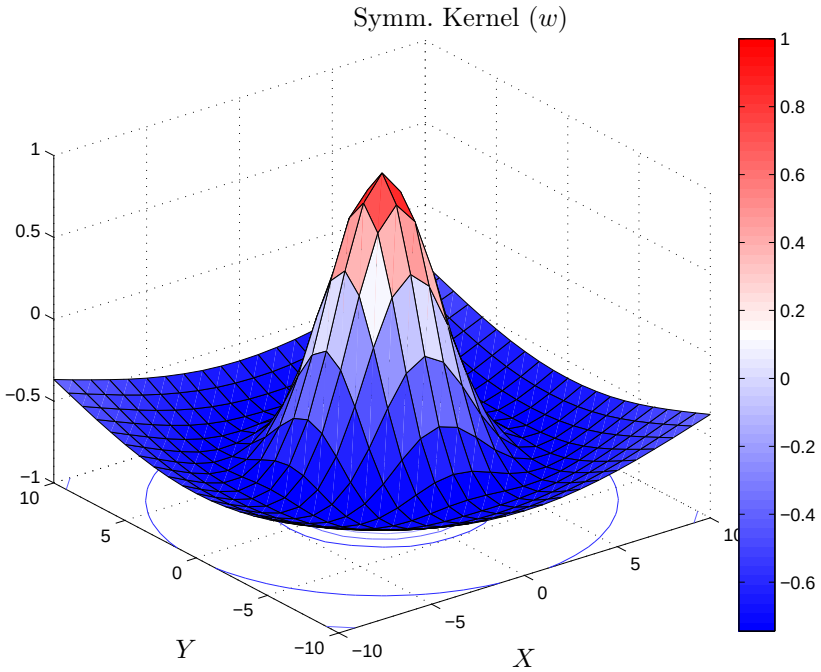


FIGURE 5.4 – Symmetric kernel in 2D. Mexican-hat with spatial support of size 21.

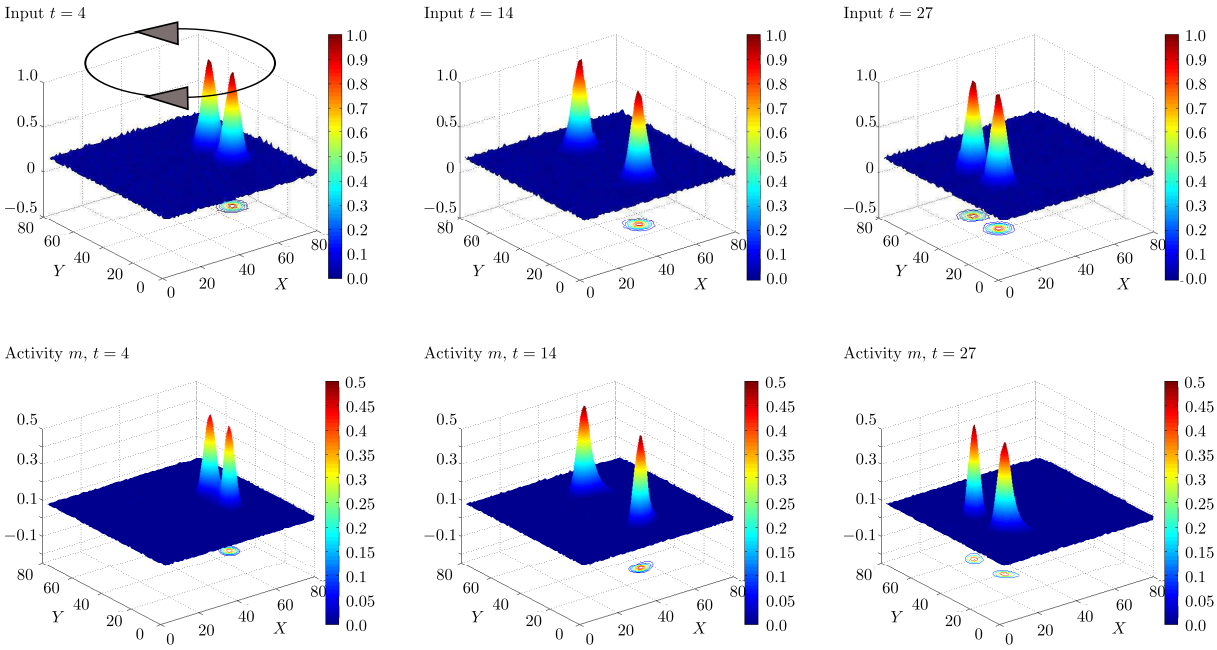


FIGURE 5.5 – Symmetric neural field dynamics. First row shows the input for the times $t = 4, 14, 27$ with a Gaussian noise of small amplitude ($\sigma = 5 \times 10^{-4}$). Second row shows the neural field activity at times $t = 4, 14, 27$ taking the first row as input. In the example two opposite stimuli generate two traveling waves with the same amplitude in the neural field. Each unit in the field has the kernel shown in Figure 5.4.

As we have explained, wave propagation can be observed in the brain and this behavior can be modeled using neural fields. In the case of a symmetric kernel the behavior of the activity of the neural population is independent of the precise input, in the sense that, if the input moves at different speeds, the neural population will behave similarly. However, slice preparations in rats have shown that this is not always the case, because asymmetries exist in the propagation of activities [PS09]. This represents a possible candidate mechanism for the coding of spatiotemporal sequences where the most prominent property is that the neural population should react differently to inputs moving at v or $-v$. To study this mechanism we introduce the asymmetry term in the kernel, to have $w(\theta' - \theta + \beta)$ in a general architecture depicted in Figure 5.6.

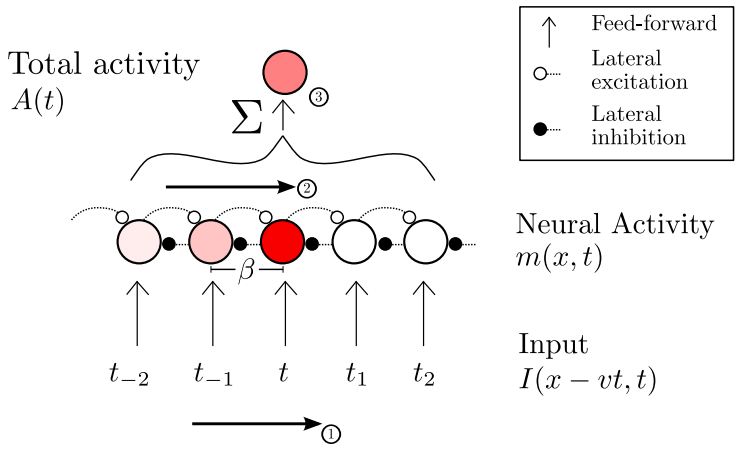


FIGURE 5.6 – CNFT with an asymmetric shape for $w(x' - x - \beta)$ (ACNFT). Given an input activity moving at speed v (1), the neural field can be tuned (shape of $w(x' - x - \beta)$) to have a higher wave amplitude for a certain input speed (2) maximizing the highest total activity in (3). In the diagram, red color intensity represents neural activity. The ACNFT also has long-range asymmetrical connections that are inhibitory (not shown)

Recalling the example of the first part of the chapter, the traveling wave amplitude of a symmetric neural field for an input moving at either v or $-v$ will be the same. In the other hand, using an asymmetric kernel generates a higher amplitude traveling wave for the encoded speed than for the inverse input ($-v$ instead of v), as shown in Figure 5.7. In this example, considering the total activity $A(t)$, and the particular input in Figure 5.7, where the input moves at the speed the system is tuned-to, the total activity is 20% higher (0.16 and 0.13 respectively, in this example), therefore, we can discriminate these two inputs directly by comparing the total activity. Yet, we need to precise the relation between β and v to be able to encode any pattern (input). We will do so by giving precise instances for the non-linear function f, the input function $I(\theta, t)$ and the kernel $w(\theta' - \theta + \beta)$.

5.2.3 Input-asymmetry function

As we mentioned, we will study the relation between the asymmetry β and the input speed v for precise instances of the kernel and the input function respectively, first in 1D and then in the 2D case.

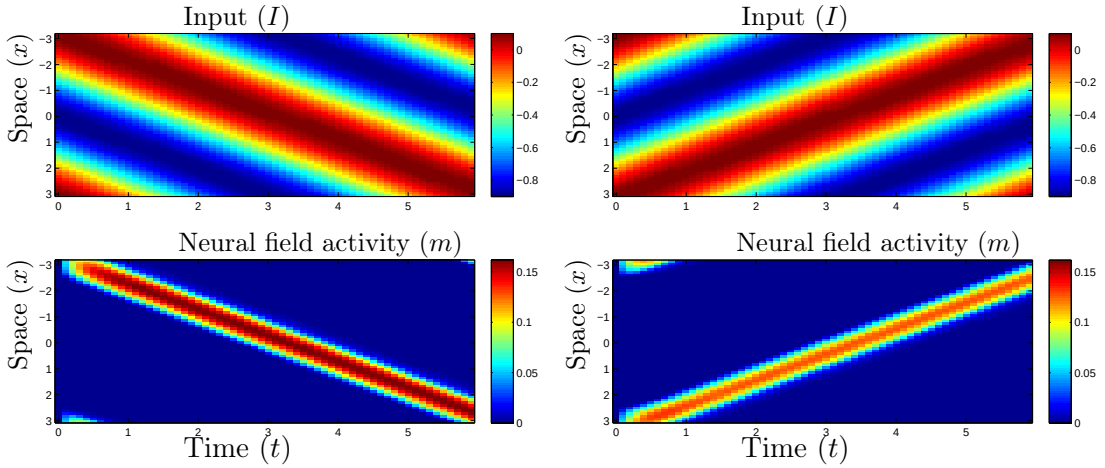


FIGURE 5.7 – (a) The input (I) bump moving at speed $v = 1$ and neural activity (m) using the asymmetric kernel. (b) Inverse input, moving at $-v$, starting at $\theta = \pi$ and its associated neural activity. Simulation parameters are $\beta = 0.13$, $\tau = 0.15$, $dt = 0.1$, $d\theta = 2\pi/60$ in the two simulations.

1D Model

The form of the input we consider is a bump to represent the activation of a precise spatial location, where we can control the ratio of maximal and minimal input, as this may influence the wave propagation. Also, it is a sinusoidal function to avoid border conditions in the integration domain,

$$I(\theta, t) = C\left[1 - \epsilon + \epsilon \cos(\theta - vt)\right] \tag{5.3}$$

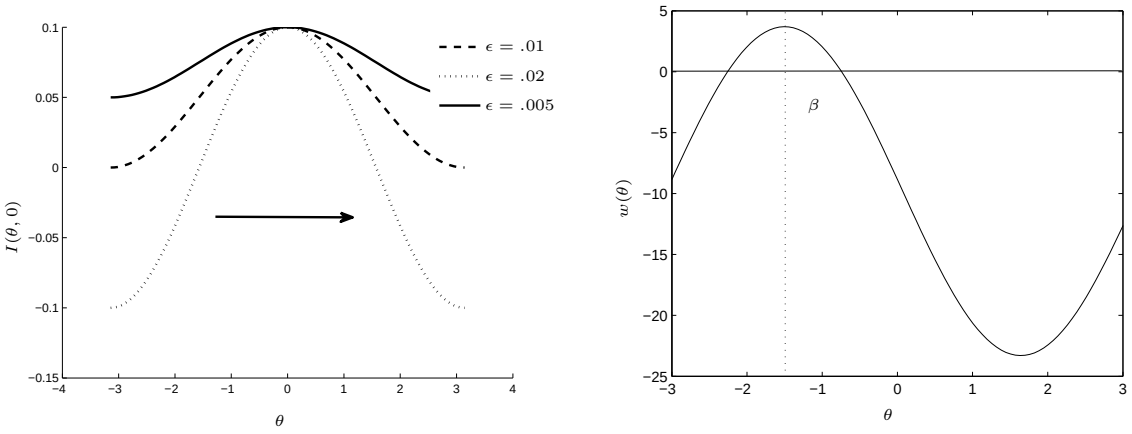
This input function can encode stimuli with different contrast levels (ratio of maximal and minimal contrast), depending on ϵ, see Figure 5.8(a). For the kernel function, what we want to obtain is a higher population response for the input moving at the speed the model is coding for (m), and a decreasing activity as the speed differs, see Figure 5.9. In particular at $-\beta$ the kernel should have its maximum and far from that point the kernel must have progressively lower values. To achieve this we consider the kernel function w:

$$w(\theta) = J_0 + J_1 \cos\left(\theta + \beta\right) \tag{5.4}$$

where the shape of w determines the coding of the system for a given speed, see Figure 5.8(b). The most prominent difference with the standard kernel functions in the CNFT is the asymmetry of this function with respect to θ. This asymmetry will make the population sensitive to one specific speed, but we need to precise this relation. Rewriting Eq. 5.5 using Eqs. 5.3 and 5.4, we obtain what we call the asymmetric CNFT (or ACNFT),

$$\frac{\partial m(\theta, t)}{\partial t} + \tau m(\theta, t) = \left[\int_{-\pi}^{\pi} w(\theta' - \theta) m(\theta', t) d\theta' + I(\theta, t)\right]^{+} \tag{5.5}$$

In Eq. 5.5 we use a linear threshold function $g(z) = [z]^{+} = max(z, 0)$ over the sum of input and weighted neighbors activities, in order to keep the activity above zero. Considering that



(a) (b)

FIGURE 5.8 – (a) Input function at different contrast levels. (b) Asymmetric kernel function for $\beta = 1.5$. Using the asymmetric kernel, we conclude that only neurons located close to a -1.5 distance from a given neuron (1.5 to the left) will have an excitatory influence.

neurons usually operate below their saturation level, the threshold applies to the integral sum and not directly to m. The model presented so far corresponds to [XG02, HS98], what follows is novel to this thesis work.

To answer if an input sequence moves with a given speed, we look at the value of the total activity over space:

$$A(t) = \int\int m(\theta, t) d\theta dt = 2\pi \int r_0(t) dt \tag{5.6}$$

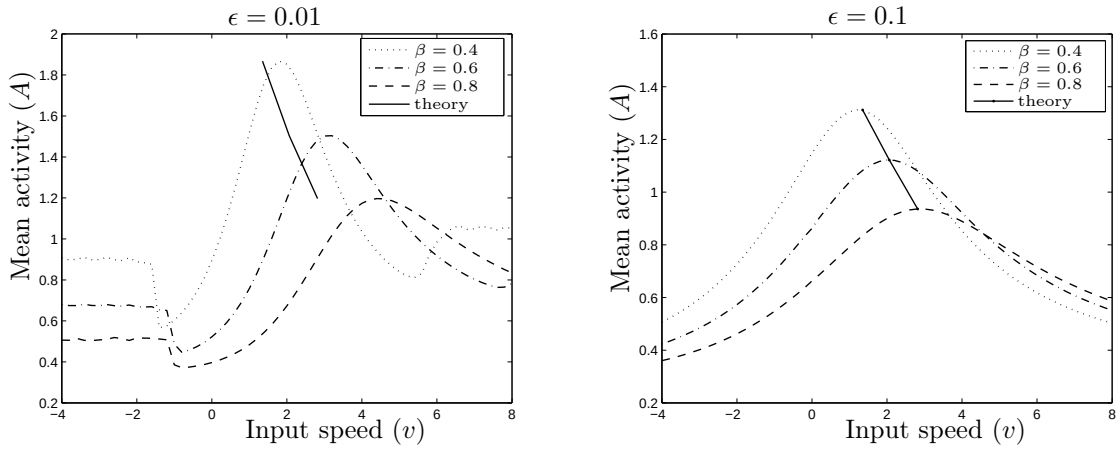
which should be maximal for a stimulus moving at the speed for which we tune the system. In Eq. 5.6, $r_0(t)$ (First Fourier component) accounts for the total instantaneous activity over space. Any other spatiotemporal sequence should deliver a lower value of A. Extending the analysis of [XG02, HS98], in this work we derive the exact expression for $v_m(\beta)$ for a fixed β, such as to maximize A (see details in Appendix A.1.2),

$$v_m = \frac{J_1 f_1(\theta_c)\tau - 2\tau\cos(\beta) - \sqrt{J_1^2 f_1^2(\theta_c)\tau^2 - 4\tau^2 J_1 f_1(\theta_c)\cos(\beta) + 4\tau^2}}{2\tau^2 \sin(\beta)} \tag{5.7}$$

all the terms in Eq. 5.7 are known except for $\theta_c(t)$, but the solution we look for must be independent of the input if we want to apply it to other sequences and link v_m with β. Considering the limit when the input width tends towards a localized input, this implies θ_c tends towards zero, simplifying v_m (see details in Appendix A.1.3),

$$v_m(\beta) = \frac{1 - \cos(\beta)}{\tau \sin(\beta)} \tag{5.8}$$

To verify this result we consider first two contrast values, and check the obtained function $v_m(\beta)$, see Figure 5.9. Then, we repeat the same analysis for several contrasts and values of β to compare against the known no-contrast limit of [XG02] (or width of the input $\epsilon \to \infty$) and our



(a) (b)

FIGURE 5.9 – Mean activity for 3 different values of the asymmetry (β). (a) Mean activity at the contrast level $\epsilon = 0.01$. (b) The same for $\epsilon = 0.1$. Continuous line corresponds to v_m computed from our high-contrast limit prediction using the 3 values of asymmetry (β).

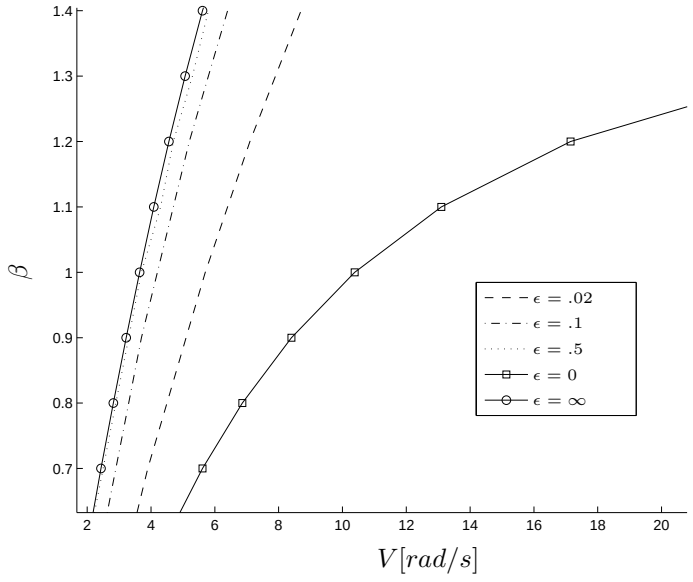
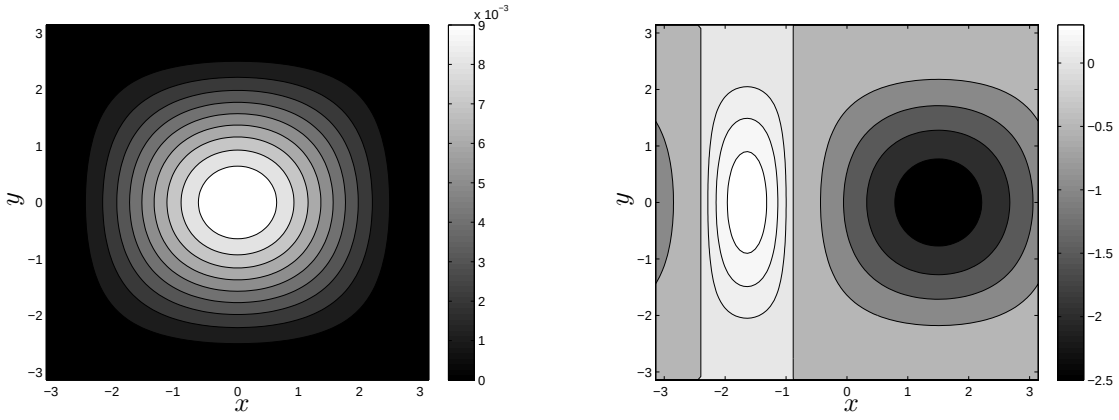


FIGURE 5.10 – The $\beta(v)$ function for the maximal mean activity (r_0) at different values of ϵ. Continuous lines correspond to theoretical results in low ($\epsilon = 0$) and high contrast limits ($\epsilon = \infty$). The other three dotted curves correspond to simulations.



(a) (b)

FIGURE 5.11 – (a) Input function. The displacement is only in the x-axis. (b) Asymmetric kernel function for $\beta = 1.5$ for a neuron at $(0, 0)$. Neurons only at $x = -1.5$ or close to that location will have positive weights.

high contrast limit at Eq. 5.8 (or width of the input $\epsilon \to 0$), and we run simulations to check that inputs with medium contrast are better represented by our high contrast limit, see Figure 5.10. It can be noticed that as the input contrast (ϵ) is higher our approximation of $v_m(\beta)$ is better, or in other terms, for a given input speed we can derive more precisely the right β so that the mean activity A is maximal. As a consequence, we can conclude that Eq. 5.8 maximizes A as function of β. Also, if we consider a Taylor expansion or $3rd$ order of Eq. 5.8 around $\beta = 0$, we can approximate the relation of β and v_m by:

$$v_m(\beta) \approx \frac{\beta}{2\tau} \tag{5.9}$$

Moreover, we notice that in the high contrast limit the relation of the optimal β and the input speed v is almost linear, as the second and third derivatives of $v_m(\beta)$ are zero at the limit $\beta \to 0$. We also check this by performing a linear regression, where we obtain a fitting with $R^2 = 0.991$ [11].

Parameter adjustment, 2D

In 5.2.3 we have introduced the model of Xie [XG02], and then we have derived an expression to link our parameter β to the input speed we want to encode in the case of 1D movements. Now we want to show that the model can be extended to encode 2D movements where speed and direction must be simultaneously encoded.

We build a classification system that consists of 2D maps of neurons. Each map is retinotopically organized, and the potential (or activity) of each neuron evolves according to the local input and to excitations and inhibitions that are received from other neurons in the framework of the neural field described in 5.2.1. These interactions are modeled by Eq. 5.10 for the activity

11. In linear regression performed by least squares (or OLS), where the idea is to minimize the distance between the data and the linear model, R^2 shows the amount of variance explained by the linear model, where values close to 1 indicate that the linear model explains most of the variance, thus the fitting has a good adjustment.

m of each unit, where we use one m (one neural map) for each pattern we want to recognize, extending our 1D model:

$$\frac{\partial m(\vec{x},t)}{\partial t} + \tau m(\vec{x},t) = \left[\int_0^{\vec{x}_f} w(\vec{x}',\vec{x}) m(\vec{x}',t) d\vec{x}' + I(\vec{x},t)\right]^+ \tag{5.10}$$

here w determines the selectivity of the system and $[]^+$ thresholds the activity to be positive-only. To study this model, we consider the case of a movement only along the horizontal axis without loss of generality, because any other trajectory can be considered as rectilinear at least locally and the kernel w can be adapted (rotated) to work in other orientations.

In 2D, the ACNFT encodes a set of trajectories in space with the same population, where each unit has a local kernel w. We consider in this analysis the case where w has the size of the input image, *i.e.* we can encode only one trajectory with the population. We notice, however, that if the kernel for each neuron has a size smaller that the neural field and trajectories are sparse "enough" (not intersecting), this will be the case in general; we will address this aspect in the discrete 2D version of our model, later in the chapter.

As in the 1D case, we consider precise instances of the kernel and input functions. The kernel function should have its maximum value in the direction of motion, precisely at $-\beta$, and a decreasing function from that point. In the direction perpendicular to the motion, the kernel should also decrease with respect to the axis of motion. Considering these elements and also to simplify the analysis in the Fourier domain, we define our kernel function w as a product of sinusoidal functions:

$$w(\vec{x},\vec{x}') = p(y,y')q(x,x') = A\left(1+\cos(y'-y)\right)\left(J_0 + J_1\cos\left(x'-x+\beta\right)\right) \tag{5.11}$$

where $\vec{x} = (x,y)$ and $\vec{x}' = (x',y')$. In Eq. 5.11 we impose $w(\vec{x},\vec{x}') = p(y,y')q(x,x')$ where the asymmetry will be only in the q function as in the 1D case, and the $p(y,y')$ function is a RBF or something close (cosine in our analysis). The separability of w helps with the analysis, since along the x-axis we have a system similar to the 1D case, see Figure 5.11(b). The input we consider is again a spatial bump moving in the x-direction without loss of generality, where we can control the ratio between maximal and minimal input, as this may influence the wave propagation, see Figure 5.11(a). The main difference with the 1D input is that this input is only positive, because the input we will study in the next chapter (related to local motion) is also positive. Considering these aspects, we take the input in Eq. 5.12.

$$I(\vec{x},t) = D^2\left[1+\cos(y)\right]\left[1+\cos(x-vt)\right] + C \tag{5.12}$$

As in the 1D case, we want to tune the parameter β of this system to maximize r_0, see Eq. 5.6. In this case several assumptions about the shape of the activity must be made, and the first objective is to obtain an expression for $r_0(t)$, see details in Appendix A.2.1. After having obtained an analytic formula for r_0, this expression is maximized, and we thus obtain in this work (see details in Appendix A.2.2) the speed that maximizes r_0 for a given β:

$$v_m = \frac{AJ_1 g_1(x_c,y_c) - 2\cos(\beta) - \sqrt{(AJ_1)^2 g_1^2(x_c,y_c) - 4AJ_1 g_1(x_c,y_c)\cos(\beta) + 4}}{2\tau\sin(\beta)} \tag{5.13}$$

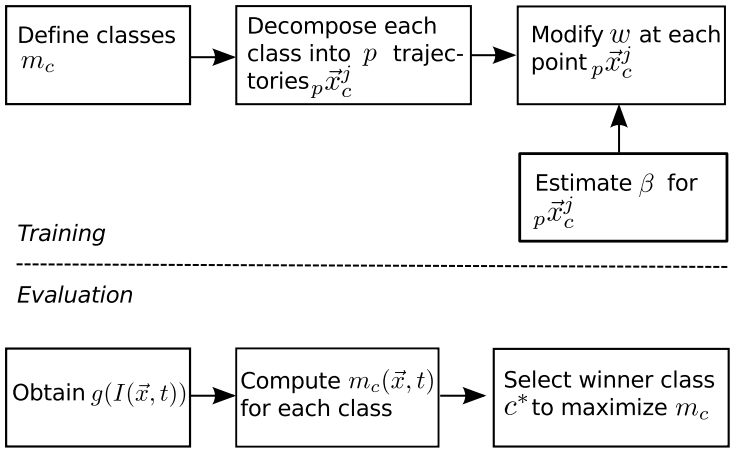


FIGURE 5.12 – Schematic view illustrating training and evaluation stages in our model for the classification task.

v_m is the magnitude of the speed, as the input is defined to be moving only in the x direction. The difference with the 1D case is the g_1 function, defined in Appendix A.2.1. For v_m we can again use the high contrast limit ($g_1(x_c, y_c) \to 0$),

$$v_m(\beta) = \frac{1 - \cos(\beta)}{\tau \sin(\beta)} \tag{5.14}$$

which is the same result as in 1D. In a similar way, the low contrast limit $v_m = \frac{\tan(\beta)}{\tau}$ can be found in the 2D case, as obtained by [XG02] in the 1D case. To verify both limits in the 2D case, we performed simulations where it can be seen that as for the 1D model at different input contrast levels (in the 2D case, the contrast is defined by C and D) the optimal value of β is close to the high contrast limit we obtained, though it is not as precise as in the 1D case. This loss of precision is probably due to several hypotheses that do not completely hold in general, see the end of Appendix A.2.1 for a discussion.

5.2.4 Discrete 2D

Until now we have considered straight-line trajectories in 2D, moving along a single axis with constant speed. Though, this is not the general case, as curved trajectories and movement with acceleration are not uncommon in patterns such as biological motion. Yet, curved trajectories can still be encoded considering them locally as rectilinear ones. As for trajectories where $dv/dt \neq 0$, they can still be encoded with our model if it is possible to consider them as piece-wise functions with locally constant speeds. More experiments should be carried out to verify both scenarios.

In the general case, if we want to encode any given trajectory we have each position available in space $\vec{x}^i = (x^i, y^i)$ in pixels and its derivative $\vec{v}^i = (v_x^i, v_y^i)$, in pixels per frame. Due to implementation issues the size of the kernel (σ) should be much smaller than the total size of the field (or image), and obviously this size limits the maximal speed the local kernel will be able to encode as $|\beta| < \sigma/2$. The minimal speed is constrained to the equivalent of 1 pixel/frame, or $|\beta| > 2\pi/\sigma$.

The pixels which are in the trajectory ($\vec{x}^i = (x^i, y^i)$) identify which kernels should be modified. Then, the direction of movement θ can be calculated as $\theta = \arctan(v_y^i / v_x^i)$ where $\theta \in [-\pi, \pi]$,

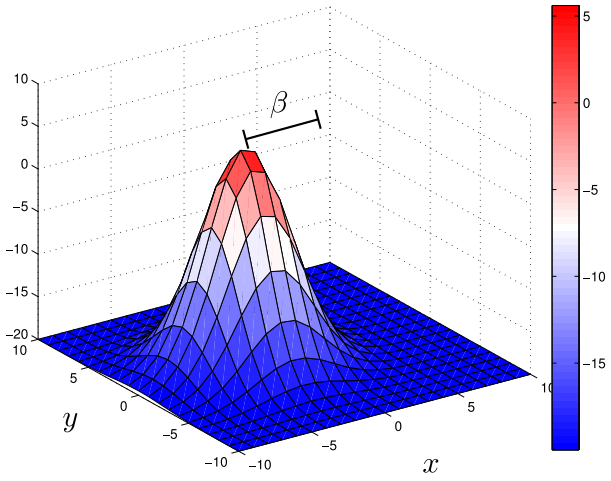


FIGURE 5.13 – Asymmetric discrete kernel for a spatial support of size 21, $J_0 = -0.2$, $J_1 = 100$ and $\beta = .8$. Note that the profile of the kernel in the plane $x - z$ ($y = 0$) is similar to the 1D kernel, where the negative amplitude is higher than the positive one. Far from this plane the kernel goes to $J_0 J_1$.

this angle is the rotation necessary to put the x-axis in the direction of motion. The magnitude of the speed in pixels $|\vec{v}^i|$ must be transformed into radians per frame with $v_r^i = |\vec{v}^i| 2\pi/\sigma$, then used so as to calculate the asymmetry in the axis of movement to obtain β^i. Considering these elements, we define the discrete 2D version of our model as,

$$\frac{\partial m(\vec{x},t)}{\partial t} + \tau m(\vec{x},t) = \left[\int_{-\pi}^{\pi} w(x_\theta(\vec{x}',\vec{x}), y_\theta(\vec{x}',\vec{x})) m(\vec{x}',t) d\vec{x}' + I(\vec{x},t) \right]^+ \tag{5.15}$$

In Eq. 5.18, x_θ, y_θ and w are defined such as to present a finite kernel, smaller than the size of the neural map, to encode more than one trajectory at a time in the same population, to represent for example, biological motion.

$$x_\theta(\vec{x}',\vec{x}) = 2\pi \left((x' - x)\cos(\theta) - (y' - y)\sin(\theta) \right) / \sigma \tag{5.16}$$

$$y_\theta(\vec{x}',\vec{x}) = 2\pi \left((x' - x)\sin(\theta) + (y' - y)\cos(\theta) \right) / \sigma \tag{5.17}$$

$$w(\vec{x}_\theta, \vec{y}_\theta) = J_1 \left\{ \exp\left(-\frac{(\vec{x}_\theta + \beta)^2 + \vec{y}_\theta^2}{2(\pi/4)^2} \right) \frac{1}{\pi^3/8} + J_0 \right\} \tag{5.18}$$

where $\vec{x} = (x, y)$ and $\vec{v} = (v_x, v_y)$ because the traveling wave is in 2D. The constant σ corresponds to the spatial support of the kernel, in general smaller than the size of the complete neural field. In these definitions, $\theta = \arctan(v_y/v_x)$ is the angle of the trajectory. Note that this expression corresponds to a translated Gaussian function, where β is the asymmetry. All the values in orientations different from θ are negative. Also, the constant J_0 makes the kernel always negative except in the zone close to $-\beta$, being the analogy to the 1D case.

Following, we look for the experimental function $v_m(\beta)$ with this new kernel in a circular trajectory, where we also modify close kernels and we compare with the bounds we computed in

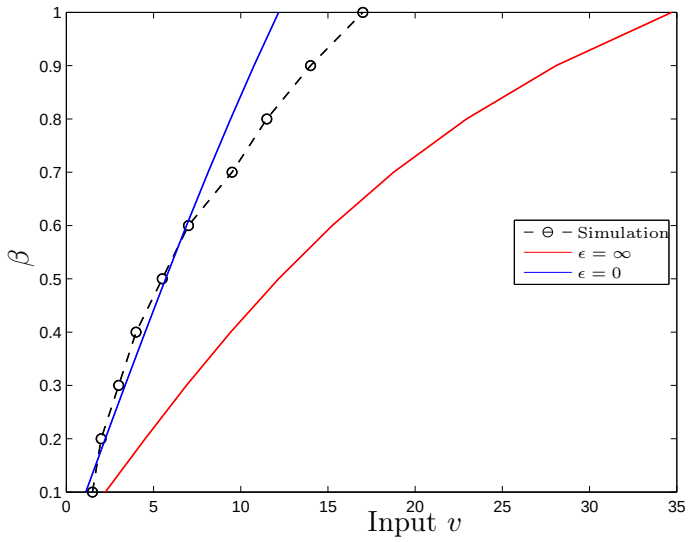


FIGURE 5.14 – Theoretical (continuous lines) and experimental functions $v_m(\beta)$ for the discrete 2D case and the kernel given by Eq. 5.18. The limit $\epsilon = 0$, corresponds to Eq. 5.8.

1D. In the discrete 2D case, as the kernel is finite ($\sigma = 21$) the bound functions must be scaled by $\sigma/2\pi$. After taking into account this factor, we can compare the fitting of the experimental versus theoretical bounds in the 2D discrete case, see Figure 5.14. It can be noticed from the comparison that our estimation in Eq. 5.8 is close to the experimental curve (blue continuous line), with a $R^2 = 0.84$. Looking more in detail, if only the interval where $\beta < \pi/4$ is selected, then the fitting improves significantly, with $R^2 = 0.97$. We can conclude that in the discrete 2D version, the asymmetric neural field follows a close to linear relation. Moreover for small β ($< \pi/4$), the fitting of the simulation and function we derive in the 1D case (Eq. 5.8) is accurate. We also notice that $\pi/4$ is the size of the Gaussian function we use in Eq. 5.18, thus the size of the interval could be related with the discrete nature of the kernel, as higher values of β will miss an important part of the positive part of the kernel.

The dynamics of the discrete system can be compared with the symmetric case that we described at the beginning of the chapter. For $\beta = 0.8$ and in a circular trajectory, the discrete version of our model generates also moving bumps, see Figure 5.15, the main difference is that amplitude will be larger for the speed the population along the trajectory is coding for.

5.3 Single Trajectories

In the following experiments, we verify the classification properties of the proposed ACNFT model for the discrete 2D case under two conditions: different speeds and different spatial trajectories, for trios of sequences. Our first observation about the discretization of the model is that the function $\beta(v)$ is still linear as we predict but it must be scaled according to the size of the kernel. We also notice that the simpler form of the kernel in Eq. 5.18 is as effective as the one in Eq. 5.11. In the following experiments, each trajectory is encoded as a moving Gaussian function with $\sigma = 3$.

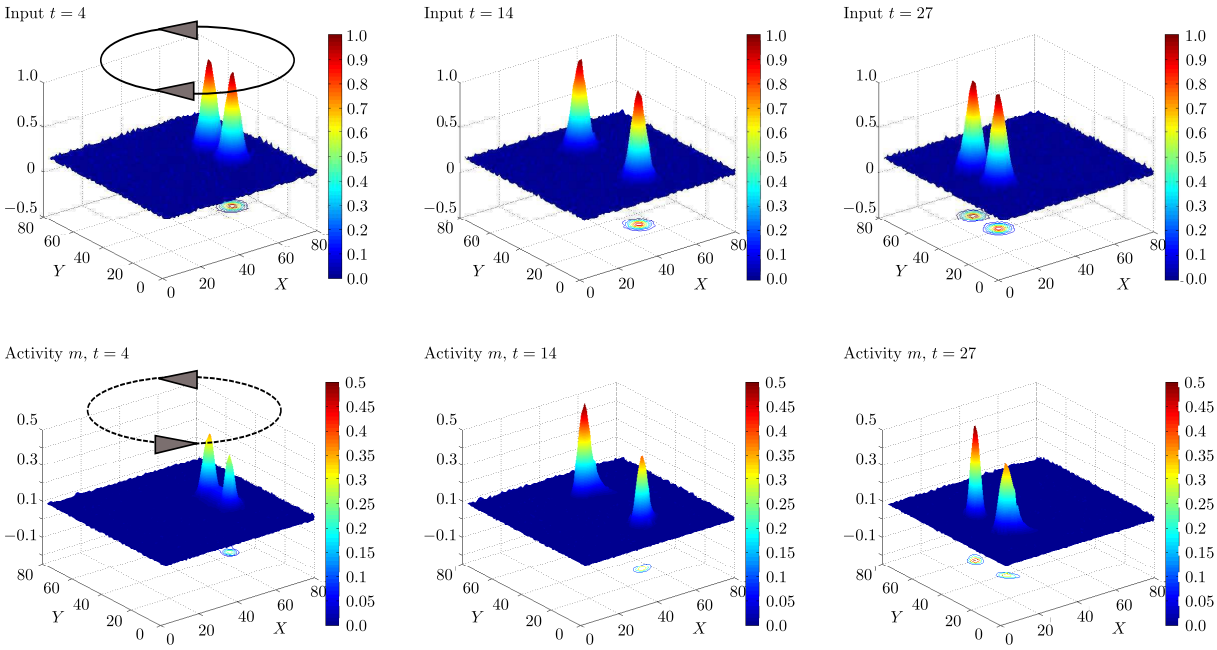
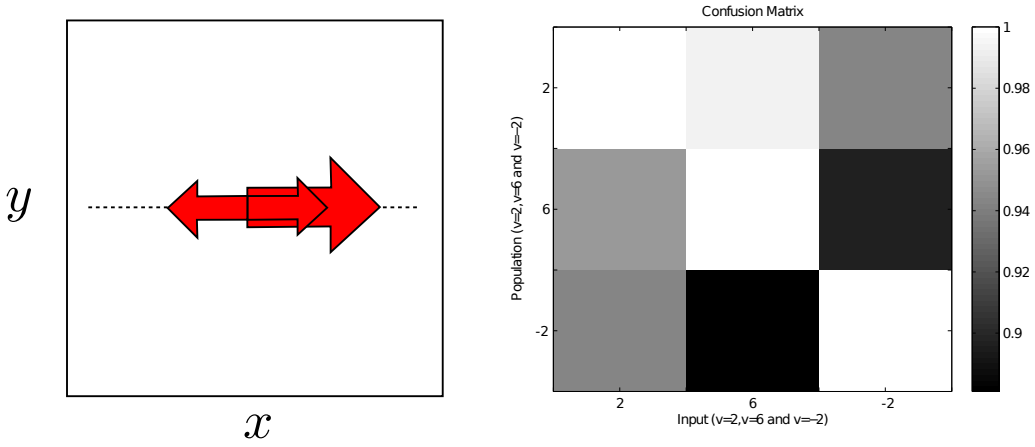


FIGURE 5.15 – Asymmetric neural field dynamics. First row shows the input for the times $t = 4, 14, 27$ with a Gaussian noise of small amplitude ($\sigma = 5 \times 10^{-4}$). Second row shows the neural field activity at times $t = 4, 14, 27$ taking the first row as input. In the example two opposite stimuli generate two traveling waves with different amplitudes in the neural field. Almost all units in the field have the kernel shown in Figure 5.4, but close to the circular trajectory the kernel is the asymmetric function given by Eq. 5.18.



(a) (b)

FIGURE 5.16 – (a) Input Trajectories. (b) Confusion matrix for 3 different speeds: $v = 2$, $v = 6$ and $v = -2$.

5.3.1 Same trajectory, different speeds

The experiment consisted in the evaluation of the discrimination of 3 stimuli: input speed $v = 2$, input speed $v = 6$ and input speed $v = -2$ all three with the same horizontal trajectory, see Figure 5.16(a). The three stimuli start from the same point with the same duration chosen, to always keep the stimuli visible. We summarize the results by first presenting the confusion matrix for the 3 patterns, in Figure 5.16(b).

The confusion matrix shows, row by row, that each population is effectively selective to a single stimulus (the one it has been tuned for). When we compare the populations for the same input (looking by columns), the highest activity is given by the right population. To present the results in more detail, we show the temporal evolutions of the total activity. We always consider the activity as $r_0(t) = \Sigma_{\vec{x}} m(t, \vec{x})/(2\pi)^2$ and the total activity as $A(t) = (2\pi)^2 \Sigma_t r_0(t)$, where m is the activity of each unit in the population.

We can see that the speed coding property in the discrete version of the ACNFT can be reproduced, where the same trajectory with different speeds, either in the same ($v = 2$ and $v = 6$, see Figures 5.17(a) and 5.17(c)) or different direction ($v = -2$, see Figure 5.17(b)) produces different total activities. The difference of two populations coding a movement in the same direction ($v = 2$ and $v = 6$) is smaller compared to when there is opposite movement, probably due to the smaller amplitude associated to higher β values. Considering the decrease of the amplitude for higher β values, this suggests that the activity should be normalized, as our earlier simulations have already shown in 5.2.3. The normalization factor can be approximated linearly as a function of β, with $R^2 = 0.98$, and a linear model defined by $m = -0.12$ and $n = 1.06$ [12].

5.3.2 Different trajectories

The experiment consisted in the evaluation of the discrimination of 3 trajectories: horizontal, diagonal and vertical, see Figure 5.18(a). The latter one starts at the superior middle corner, the

12. In this case, we note linear model parameters as $v = m\beta + n$.

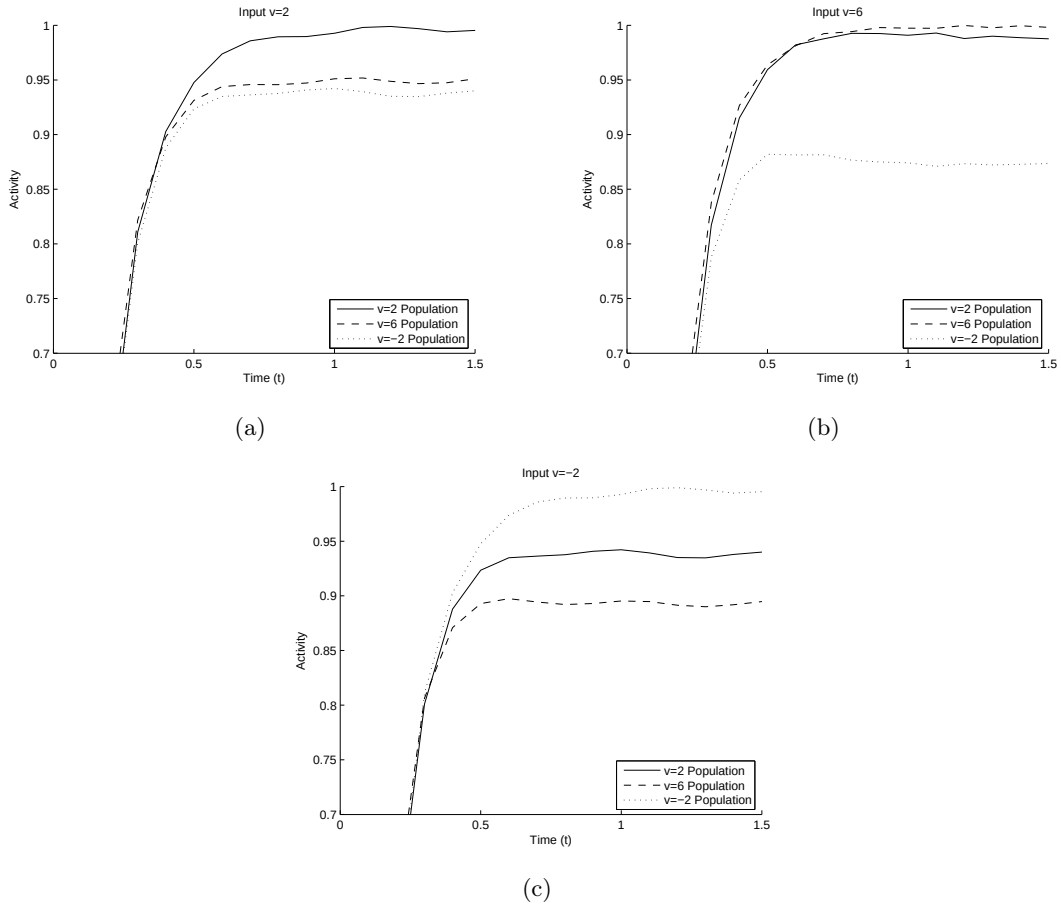
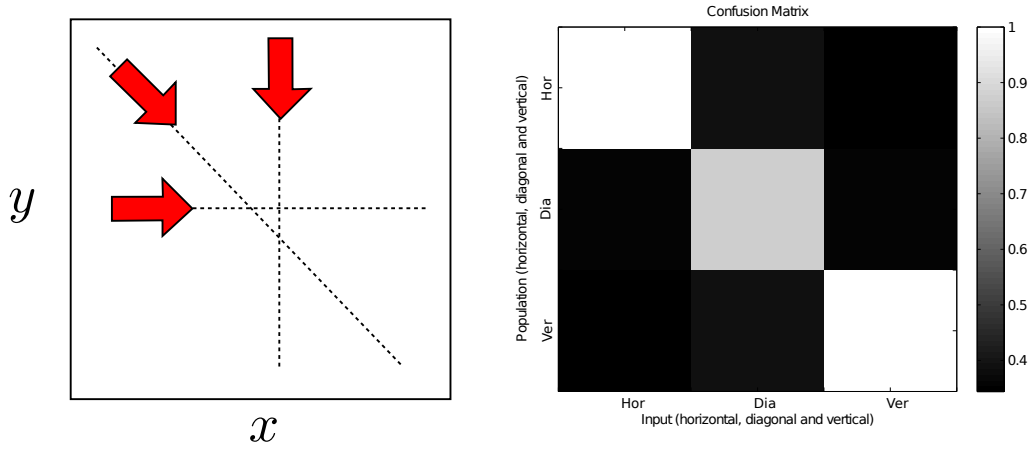


FIGURE 5.17 – Temporal evolution of the activity r_0 for the 3 populations encoding: $v = 2$, $v = 6$ and $v = -2$. Here the input moves at speeds (a) $v = 2$, (b) $v = 6$ and (c) $v = -2$.



(a) (b)

FIGURE 5.18 – (a) Input Trajectories. (b) Confusion matrix for 3 different trajectories: horizontal, diagonal, and vertical (white pixels represent higher values).

second at the superior left corner and the first one in the center left position. We summarize the results by first presenting the confusion matrix for the 3 patterns in Figure 5.18(b).

The confusion matrix in Figure 5.18(b) shows, row by row, that each population is effectively selective to a single stimulus, since its activation is higher for the stimulus it has been tuned for. At the same time, when we compare the populations for the same input (looking by columns), the population that presents the highest activation is the one that has been tuned for this input. To present the results in more detail we show the temporal evolutions of the total activity for the horizontal, diagonal and vertical input in Figure 5.19(a), Figure 5.19(b) and Figure 5.19(c) respectively. Here we always consider the activity as $r_0(t) = \Sigma_{\vec{x}} m(t, \vec{x})/(2\pi)^2$ and the total activity as $A(t) = (2\pi)^2 \Sigma_t r_0(t)$, where m is the activity of each unit in the population.

As we can see in Figures 5.19(a), 5.19(b) and 5.19(c), differences in the trajectories generate important differences in the total activity, sufficient to already provide a good class discrimination. Of course, this can be achieved even with symmetric kernels, because only the neurons in the trajectories are concerned, but we were interested in understanding discretization effects and how the model reacts to intersections in the trajectories. About discretization effects, horizontal and vertical trajectories respond as predicted but the diagonal one shows a lower amplitude than the horizontal or vertical case and small amplitude oscillations, see Figure 5.19(b). Probably this effect is due to the effect of writing the rotated version of the kernel in a small discrete patch, where discontinuities may appear. About intersections, there is a time interval when the intersecting trajectories have the same amplitude, see Figure 5.19(a) at $t = 30$. We conclude that if the intersection between two trajectories is too large the discrimination will be much more difficult. Compared with speed variation, different trajectories produce higher variations of the total activity: speed variation changes the relative difference among populations from 5% to 12%. On the other hand different trajectories can generate a difference of up-to five times, in the total activity.

In this set of experiments, we conclude that our proposed model is able to encode both different speeds and trajectories, in the discrete implementation. At the same time, it was verified that our previous theoretical analysis was generic, because the model tolerates short trajectories intersection and encodes trajectories that are not horizontal.

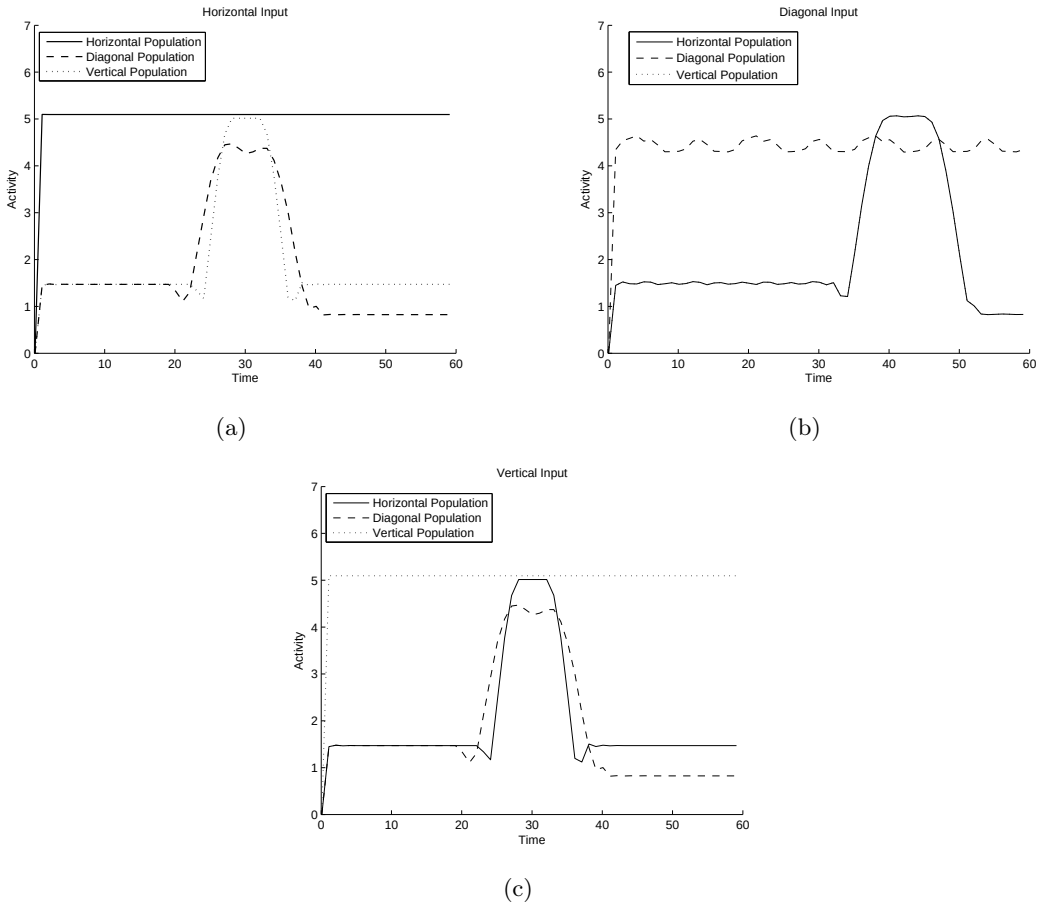


FIGURE 5.19 – Temporal evolution of the activity r_0 for the 3 populations: horizontal, diagonal, and vertical. The input is (a) horizontal, (b) diagonal and (c) vertical.

5.4 Discussion

In this chapter, we present our model for the encoding of a set of spatiotemporal trajectories by means of a neural field with the specificity of having an asymmetric kernel to bias the activity in a particular direction. The results are both theoretical and experimental and show the effectiveness of this approach. The model supposes that the pattern to be encoded can be broken down into a set of local trajectories that can be short, and do not necessarily cover the complete pattern (in space or time). We also show, by means of simulations using simple trajectories, that when several trajectories are encoded by the same map (or several inputs), the overlapping among trajectories reduces the performance of discrimination. However, classification still remains possible if the intersection is not complete.

The ACNFT proposes a way that spatiotemporal sequences encoding could be performed in the brain, not by explicitly encoding 2D snapshots of the visual scene but by implicitly recording the local features that describe the global pattern. Our model is not necessarily saying that human recognition of visual patterns could entirely rely on local patterns; experiences such as the random PL stimuli have shown that this interpretation does not have a strong support. However, our model hypothesizes another possibility: local patterns may be derived from an internal representation of the body to recognize them, like an internal "simulator" (or from some shape cues).

The proposed model illustrates how a population of units, the neural field, can encode a global pattern by means of a distributed representation. The only constraint is that the pattern must be decomposable into local trajectories. Our distributed model for pattern classification illustrates that not only "simple" tasks such as motion detection, can be addressed in a distributed approach. High level tasks, such as classification of temporal sequences, can be solved using a similar framework. In the next chapter, we will apply our model to real data to verify if complex human motion sequences can be decomposed into local trajectories, and at the same time, if our model is able to encode them and reproduce properties of human visual sequence-discrimination capabilities. In particular, we will explore the case when $dv/dt \neq 0$ and short trajectories, which are very common in biological motion. Additionally, using our model as discriminating metric, we will compare two features to be used as input: raw motion detection and local optical flow operators to determine the functional advantage of including intermediate features like local flow operators in the recognition of visual sequences.

6 Evaluation

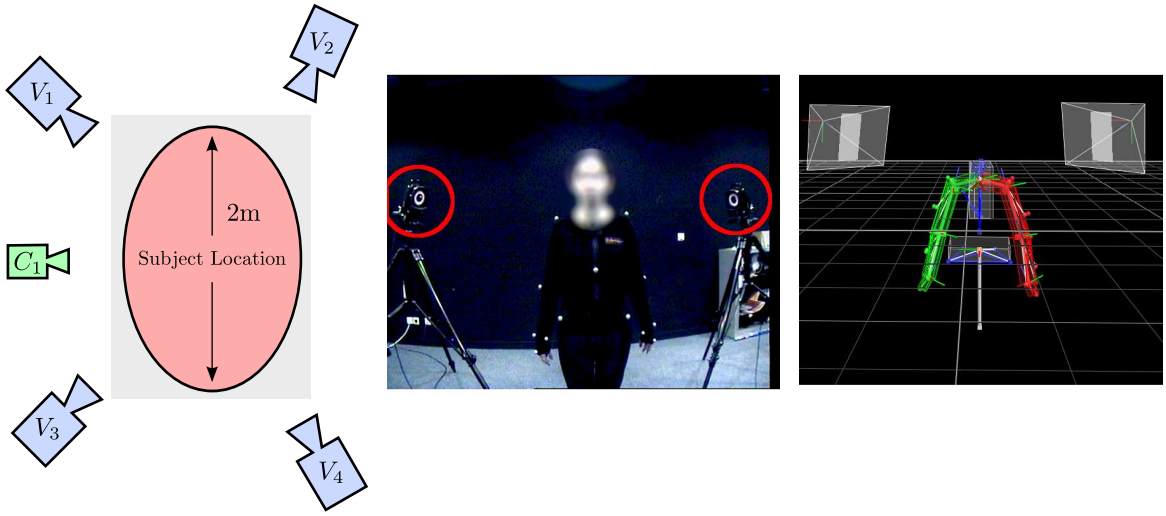
Contents

The previous chapter presents a model (ACNFT) for the discrimination of visual sequences, based on the continuum neural field theory. The model is able to encode a set of spatiotemporal trajectories in a distributed framework. In this framework, a population of units represents a spatiotemporal sequence. For our ACNFT model, we have shown theoretically and experimentally that in the case of trajectories with uniform speeds and with a perfect bump input, the model can satisfactorily encode trajectories and discriminate among them. Yet, we still have to show that the model can handle non-uniform trajectories and noisy inputs due to local speed detection.

In the first part of this chapter, we evaluate the performance of the ACNFT model using realistic local trajectories captured from a set of cameras. We verify the model and study its properties under more realistic conditions. Later, we evaluate the performance of our model using two possible feature extraction methods: (1) the raw optical flow and (2) the decomposition into local flow patterns. The main objective of the chapter is to confront our proposed model with real motion sequences, captured in 3D and 2D, and discern the properties of the human action recognition we can retrieve.

6.1 Sequence discrimination

In this Section, we study a database of three movements: “to clap”, “to fight” and “to wave” and analyze the discrimination achieved using our proposed ACNFT model. The database was



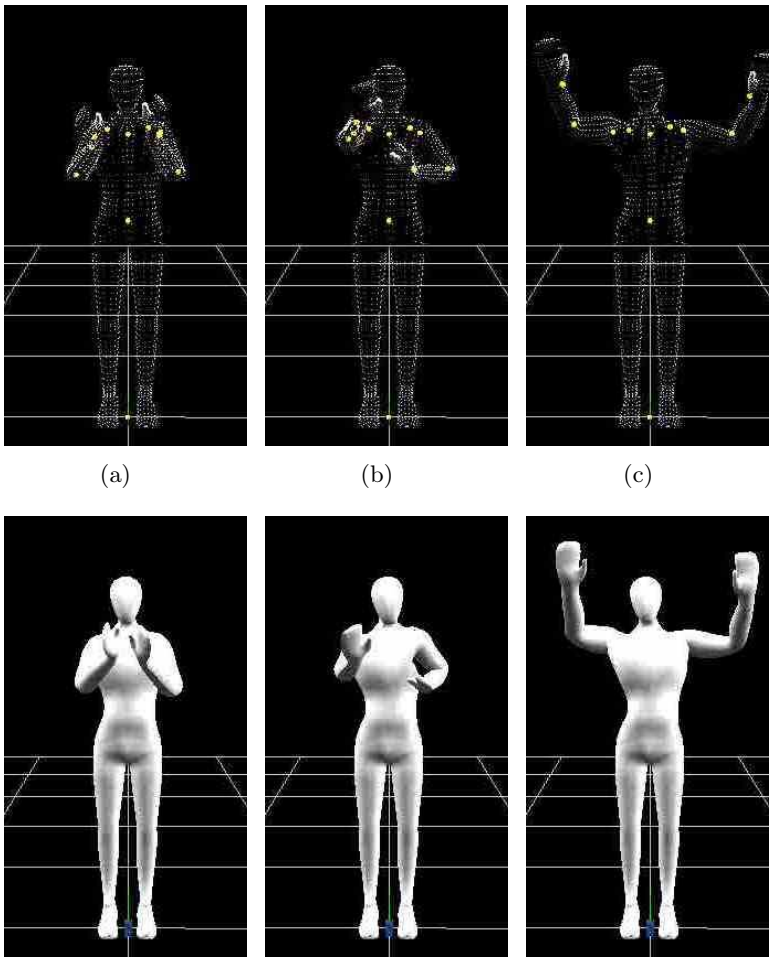
(a) (b) (c)

FIGURE 6.1 – (a) The acquisition setup, V_1 to V_4 are the VICON© cameras, C_1 is a normal 25Hz reference camera. (b) Subject DA is in the rest position facing camera C_1. Cameras V_2 (left) and V_4 (right) are indicated by the red circles. The subject is marked with 2.5cm reflective markers. (c) The result of the acquisition showing markers and joints.

created during this thesis work using a VICON© system, where reflective markers are placed on the body of a subject. A set of cameras (4 in the available system) reconstructs the 3D position of each marker, see the acquisition setup in Figure 6.1(a). After the 3D position of these markers has been obtained, a kinematic model [KRW90] is adapted to the anatomical measurements of the subject (leg length, hip size, height). Taking the 3D marker positions and the kinematic model into account, the VICON© system can directly retrieve the position and orientation of each joint as in Figure 6.1(c). To correctly determine the position and the orientation, at least 3 markers must be placed on each segment (bone).

The captured database contains the 3D coordinates of markers and joints at each time instant. It also gives the orientation of the associated bone (segment) for each joint. One of the advantages of this movement database is that it can be visualized either as just points (like the point-light stimuli) or with the complete body (using a computer body animation technique). The sampling rate of the capture was 100 Hz (100 positions per second), allowing to subsample the movement at lower frequencies, in contrast with the available standard video databases commonly available only at low frequencies. We can also observe the sequence from any viewpoint in the 3D volume as the capture is in 3D. The higher sampling rate and the capacity to observe the movement from any point of view are the main reasons to perform our own data acquisition. Public databases, like KTH (see Appendix C.2), record sequences from a few view points at standard video recording rates (25-30 Hz). The spatial error of the setup is around 2 mm for each marker. Our database contains information of one female subject (DA) who has normal athletic skills and accepted to participate in the experience, see Figure 6.1(b).

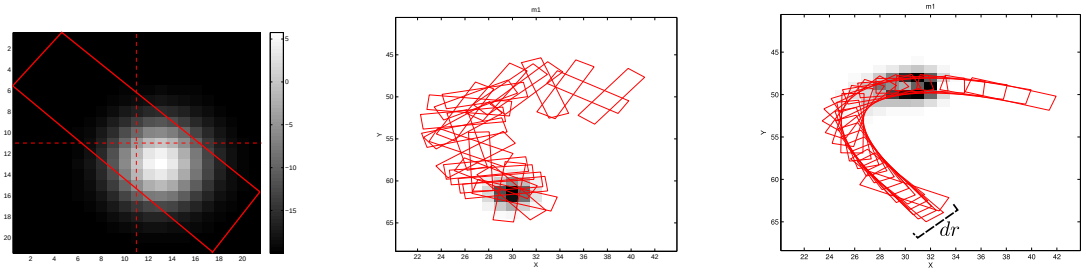
As we mentioned before, the information can be visualized either as a PL stimuli, or as an animation. The PL stimuli can be obtained straightforwardly from the data and a model of a



(a) (b) (c)

(d) (e) (f)

FIGURE 6.2 – The movements we study: “to clap”, “to fight” and to “to wave” (by rows). The images (a), (b) and (c) show only the joints (PL stimuli). The images (d), (e), (f) are the animated results, combining joints, body mesh and skeleton information.



(a) (b) (c)

FIGURE 6.3 – (a) Representation of the oriented kernel. (b) The kernel disposition in random configuration for the "to clap" sequence. (c) The ACNFT kernel configuration for the "to clap" sequence.

camera to project the 3D points onto the 2D camera plane (see more details in Appendix B). An example frame from each sequence using the PL stimuli can be seen in Figures 6.2(a), 6.2(b) and 6.2(c).

The second kind of stimuli we generated with the data was body animations. Animations were performed using a standard skeleton based animation algorithm implemented in the 3D Graphics Engine OGRE [Jun06]. An animated frame for each movement can be seen in Figures 6.2(d), 6.2(e) and 6.2(f). In skeleton animation, a static mesh represents the surface of the body and at the same time this mesh has an associated skeleton, so that each vertex in the mesh can be controlled solely from the joint information. To perform this control, the skeleton associated with the body mesh must have a direct and unique relation with the captured 3D joint data, i.e. for each 3D joint there is a single corresponding joint in the skeleton to allow the deformation of the mesh. Then, to deform the mesh, the information of several (up to 3) joints is considered (for more details see [Jun06]). The mesh model we use was originally available from the MotionBuilder© clip art and made public by [Lee11]. The skeleton was rebuilt to match the kinematic model used in the capture pipeline. The animation software was originally developed by [Lee11] but extensively modified to be used in this thesis, in order to make it compatible with the VICON© system output format.

6.1.1 Classification using the ACNFT model

In Chapter 5 we have proposed a model for the discrimination of sequences. In this chapter we perform the capture of 3 sequences, to verify the discrimination properties of our model. In order to perform this, we build 3 populations: m_1, m_2 and m_3 as defined in Eq. 5.10, one for each sequence, and we observe the total activity $A(T) = 2\pi\Sigma_t\Sigma_{\vec{x}}m(t,\vec{x})$ where T is the total length of the sequence. In the next set of experiments we use only the trajectory of the left wrist to train each model. The kernel w is modified at each location $\vec{x}$ where the wrist goes through and in the close neighborhood defined by a parameter $dr = 1$, see Figure 6.3(c). When $dr = 0$, it corresponds to modifying the kernels only along the wrist trajectory. As dr increases, the kernels of neurons close to the wrist trajectory are modified also. For example, $dr = 1$ corresponds to modifying one neuron's kernel to the left and to the right of the trajectory in the direction of the movement. The shape of the kernel at each location is set according to the wrist velocity vector using the linear relation between the speed and β derived in Chapter 5.

TABLE 6.1 – Confusion matrix for the classification. Populations are by rows, and different inputs by columns

	To clap	To fight	To wave
To clap (m_1)	34.20	5.63	5.60
To fight (m_2)	5.62	21.19	5.60
To wave (m_3)	5.62	5.63	24.21

To check the discrimination of the ACNFT model we measure $A(t)$ considering the 3 populations and using as input the three possible wrist trajectories. The results are shown in Table 6.1 where we can see that the highest activity is always in the diagonal, showing that each population is effectively coding one movement. It can also be noticed that outside the diagonal, the activity is not zero because in Eq. 5.10 even if the kernel does not match any trajectory there is anyway a (low) activation. From the same Table 6.1 we notice that the trajectory "to fight" seems more difficult to represent as the absolute difference between the wrong and right input (second row) is weaker.

The next experiment aims at verifying if each population is correctly coding the trajectory, and also the effect of time compression/dilatation. To verify the behavior of the model under these perturbations, we consider each population (m_i) and the precise optimal input for that population, for example the "to clap" input for the population coding "to clap". In this setup, we change the speed-up factor (S_u). When $S_u = .5$ it means that the input is presented at half of the normal speed. Our experiment searches to verify if the asymmetry of our kernel function w induces a larger total activity than a random kernel orientation, in the same conditions (kernel size and amplitude), so that we are effectively coding the presented trajectory even if the input speed is increased or reduced. In this experiment, the trajectories are not uniform as the ones considered in Chapter 5, hence much more difficult to represent as the speed is not constant.

For each input, we consider the activity A using our asymmetric kernel (ACNFT), see Figure 6.3(c), and we compare it with the mean activity from a (uniformly) random orientation (<Random>), see Figure 6.3(b). To check the velocity coding we vary the speed-up factor (S_u) and we repeat the experiment 100 times to avoid any bias in the orientation. In each of these runs we try a random orientation for each neuron in the trajectory (uncorrelated). To compare the value using the ACNFT kernel and the mean of the kernel we choose to perform a one-sample and paired-sample t-test [Sap06], to verify that the ACNFT activity is higher than the mean of the random kernel model, and that this difference is statistically significant ($p < 0.05$).

The t-test is one of the statistical tests that can be used to verify if a given constant is equal (or not) to the mean of a certain normally distributed series. The one-sample variation has a similar statistical hypothesis, that verifies if a given value is greater (or smaller) than a certain constant rather than equal, again normality of the series is assumed. In our case the series is the set of 100 repetitions of the random orientations of the kernel, and the constant is the mean activity of the population using the ACNFT model. As the orientation is determined by a uniform random distribution, then the values follow typically a normal distribution around the trajectory orientation.

The results of this experiment for the "to clap" (Table 6.2), the "to fight" (Table 6.3) and the "to wave" (Table 6.4) sequences show that the activity of the ACNFT is significantly higher than the random orientation ($p < 0.05$) in almost all tested cases. Furthermore, that the absolute difference between the mean activity of the random and the ACNFT activity (ACNFT - <Random>) decreases when the speed-up parameter (S_u) increases. The total activity A (ACNFT in

Table 6.2 – T-test Population 1, "To clap"

Speed-Up (S_u)	<Random>	ACNFT	ACNFT - <Random>
0.5	22.74	23.56	0.82
0.6	21.33	22.03	0.70
0.7	20.61	21.21	0.60
0.8	19.20	19.69	0.49
0.9	17.86	18.30	0.44
1.0	16.75	17.06	0.31
1.1	16.26	16.55	0.29
1.2	15.34	15.61	0.27
1.3	14.88	15.07	0.20
1.4	13.97	14.15	0.19
1.5	13.69	13.84	0.16

Table 6.3 – T-test Population 2, "To fight"

Speed-Up (S_u)	<Random>	ACNFT	ACNFT - <Random>
0.5	19.46	19.97	0.52
0.6	18.55	18.97	0.41
0.7	17.06	17.46	0.40
0.8	16.59	16.86	0.27
0.9	15.39	15.64	0.25
1.0	14.83	15.02	0.19
1.1	14.28	14.46	0.17
1.2	13.53	13.64	0.11
1.3	12.93	13.01	0.08
1.4	12.51	12.62	0.10
1.5	12.29	12.35	0.06

tables 6.2, 6.3 and 6.4) is always reduced when the speed-up (S_u) increases because consecutive inputs are closer, inversely when S_u decreases the considered inputs are more separated (in all S_u, the same number of iterations is considered). However, we also expected a reduction of the ACNFT total activity when S_u goes below 1 as predicted in the previous chapter (higher total activity for a certain β, decreasing for higher or lower speeds), but in fact for irregular trajectories like the ones tested ($dv/dt \neq 0$) the system does not have this property. We also notice the last values of Table 6.4, where the <Random> population, is higher than the ACNFT (also $p < 0.05$). In this sequence (to wave), the movement is particularly rapid, and when $S_u > 1.2$, then the movement is faster than the speed encoded by the highest considered value of the asymmetry in our implementation ($\beta = 0.8$), thus the ACNFT starts to miss some of the points along the trajectory.

6.1.2 Model Properties discussion

We now consider some of the properties that were enunciated as observed in the human visual perception of sequences and we study if they can be retrieved using our proposed model with the database we built: timing, view dependency, and incremental discrimination.

TABLE 6.4 – T-test Population 3, "To wave"

Speed-Up (S_u)	<Random>	ACNFT	ACNFT - <Random>
0.5	16.96	17.31	0.35
0.6	16.37	16.67	0.31
0.7	15.33	15.49	0.16
0.8	14.70	14.89	0.19
0.9	14.06	14.18	0.11
1.0	13.23	13.27	0.04
1.1	12.62	12.64	0.02
1.2	12.77	12.75	-0.02
1.3	12.48	12.53	0.05
1.4	11.90	11.86	-0.04
1.5	11.31	11.17	-0.13

Rapid discrimination

Similar results as in Table 6.1 can be observed over time, in Figure 6.4. In each test, the population mean activity is significantly different from the other two populations. To compare the total mean activity, we choose to use the One-Way ANOVA test ($p < 0.05$) [Sap06]. This test compares two data series means, where the series represents independent samples containing mutually independent observations, issued from normal distributions. In our experiment, each series is defined by the activity at each time t for each population that we consider normal because usually the activity for each populations moves around a certain fixed value. The value p indicates that the probability to affirm that the two series (populations mean activities) are different, when they are equal, is less than 5%. To compare more than two series (in this case three), we use a multiple test similar to the one-way ANOVA. This multiple test is necessary because when two (or more) sequential one-way ANOVA are performed, the probability of error will increase with the number of comparisons, but what we want to verify is if any of the series are similar.

Considering the multiple test, the other two populations do not significantly differ from each other ($p < 0.05$). Even more, both curves overlap in Figure 6.4(a), for example. It can be noticed that the temporal activities differ very rapidly, more precisely, they appear significantly different from the second iteration. Thus the discrimination can be achieved very quickly. However, the activity is not constant, as we consider in the analysis of Chapter 5 for A (or r_0), because the movements we consider do not have constant speeds. Regardless of the different speeds, the coding of the ACNFT is effective to encode movement even with irregular speeds, allowing very rapid discrimination at the same time.

View dependent discrimination

View dependency is characteristic of visual sequence discrimination. In this experiment we want to show how a rotation in the plane of the camera, see Figure 6.5(a), affects the discrimination of one population. We only consider the "to clap" population, and rotate it in $0\,°, 5\,°, 10\,°, 15\,°, 20\,°$ around the camera for three possible inputs, "to clap", "to fight" and "to wave". The results can be observed in Figure 6.5(b), where for each rotation the three inputs are used, but one population is considered (m_1). It can be also noticed that the mean activity is larger for the "to clap" population than for the other inputs at 0° and 5° degrees of rotation,

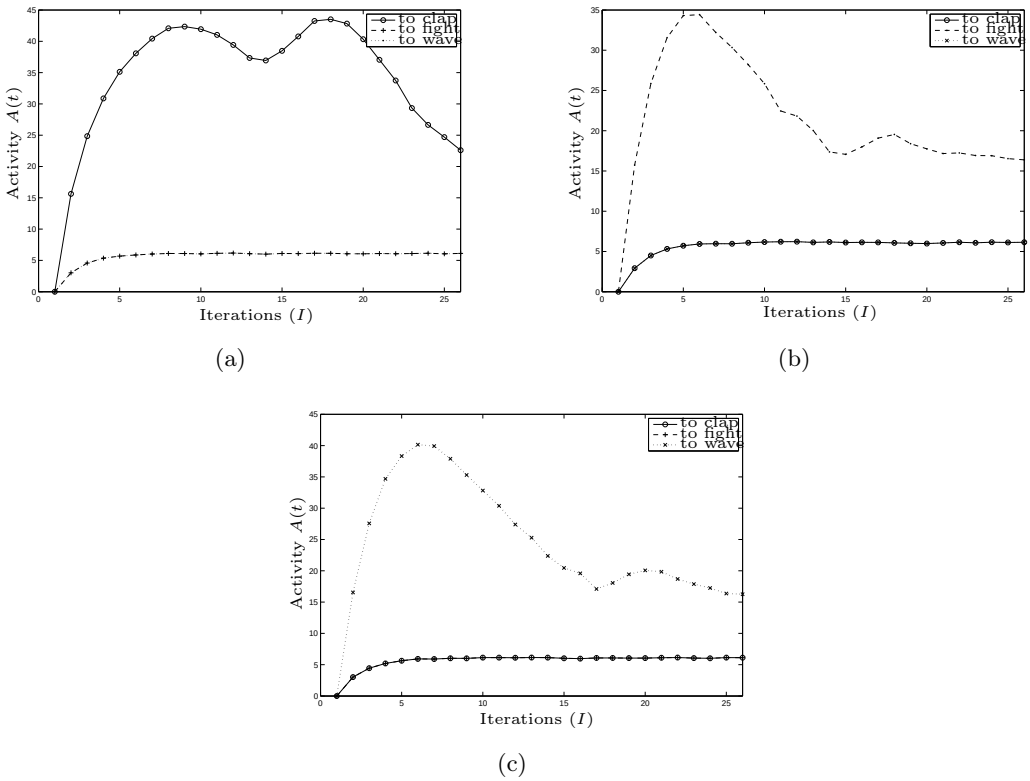
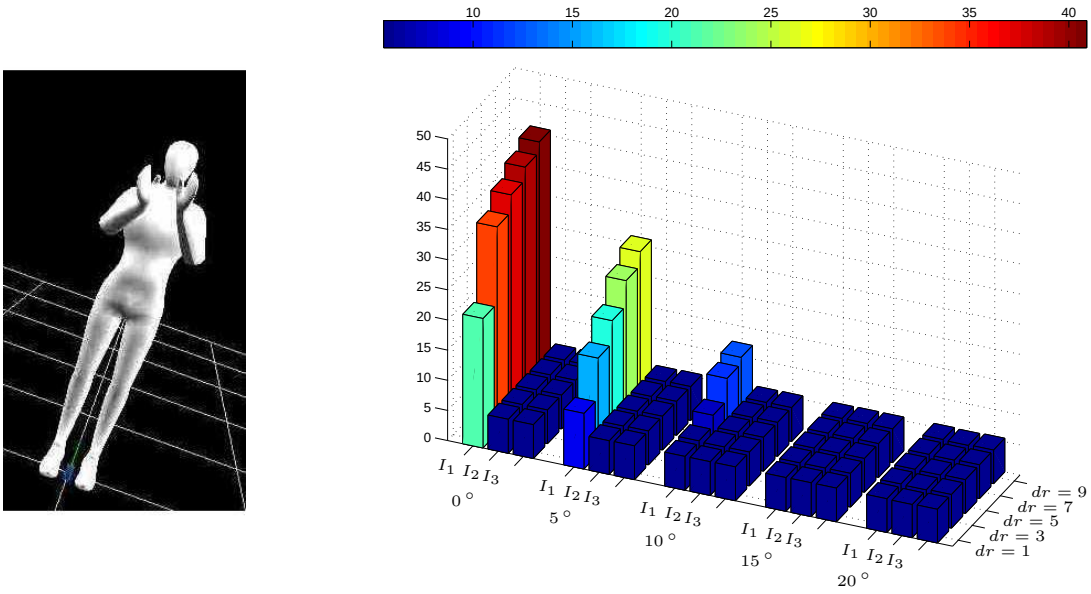


FIGURE 6.4 – Activity of each population for the different inputs. (a) The m_1 population ("to clap") activity as function of iterations for the three possible inputs. The highest activity corresponds to the "to clap" input. (b) and (c) the same for the m_2 population ("to fight") and the m_3 population ("to wave"), respectively.



(a) (b)

FIGURE 6.5 – (a) A frame of the "to clap" sequence rotated by 20 °. (b) Mean activity of the "to clap" population for the 3 possible inputs. The population is always trained with the original sequence (0 degree rotation). Each group shows the mean activity for three inputs: I_1, I_2 and I_3, respectively "to clap", "to fight" and "to wave". Also the effect of a change in the dr parameter can be observed. Higher values of dr increase the tolerance to rotations.

then the stimuli become indistinguishable. We also consider the change in dr (the size of the neighborhood around the trajectory where kernels are modified), from the default value $dr = 1$ up to $dr = 9$ (taking only odd values, to avoid discretization problems in the kernel). The effect of this variation is to increase the discrimination rate.

Incremental (additive) discrimination

One interesting aspect of the classification is to check if several (or longer) trajectories can deliver better discrimination. In Figure 6.6 (one joint) we can observe the mean activity, for the "to clap", "to fight" and "to wave" populations, for all the inputs, as we have seen in Table 6.1, for example the "to clap" population (m_1) presents a higher activity for the input that corresponds to what the population encodes for I_1. We extend this analysis considering two joints, see Figure 6.6, in this case we can notice that the total activity is higher when we consider 2 joints, close to twice as high. Yet, at the same time the minimal activity also doubles. By consequence, the discrimination rate (proportion between minimal and maximal activity) increases as the number of considered trajectories increases. However, as the minimal activity also increases when considering several trajectories absolute comparisons of the mean activity among populations with different numbers of trajectories should be carefully performed.

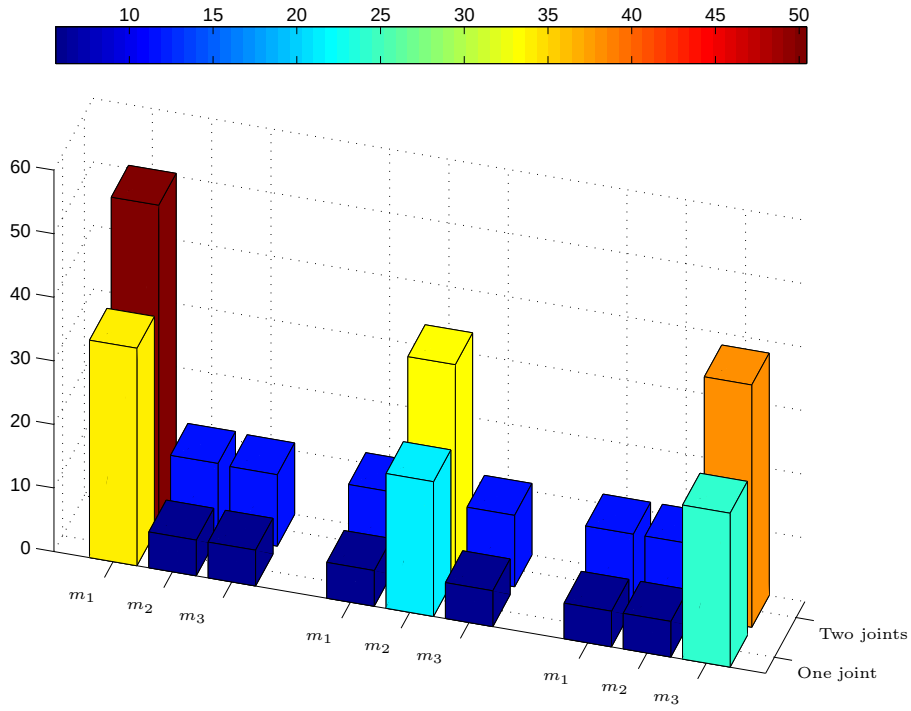
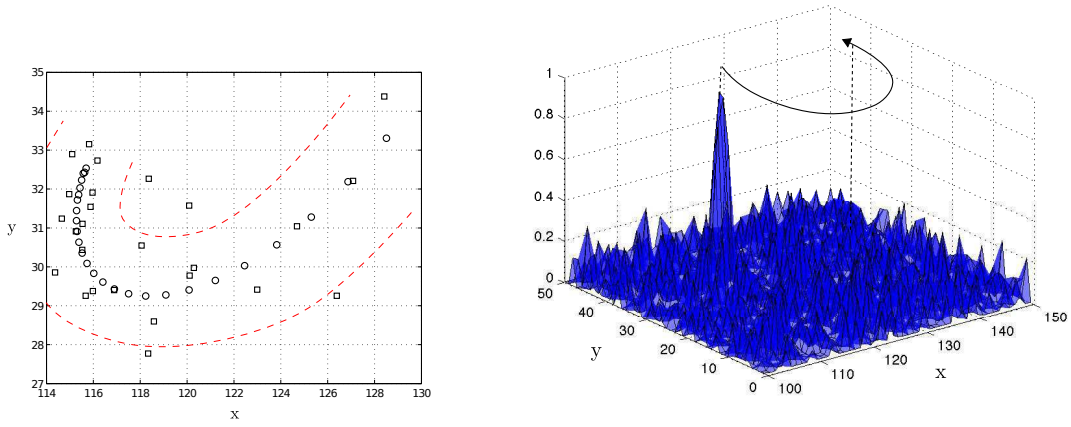


FIGURE 6.6 – Mean activity of each population when 3 possible inputs are applied. In the first group (input "to clap"), the highest activity corresponds to population m_1 because the input corresponds to this sequence (m_1 encodes "to clap"). The second test applied to the group corresponds to the case of two joints. Here a higher activity than in the one joint case is observed.



(a) (b)

FIGURE 6.7 – The two inputs to study the generalization properties of the ACNFT model. (a) Trajectory deformation, with $\rho = 1$. Circles represent the encoded trajectory, and the squares represent the input position. (b) Additive noise, with $\nu = 0.01$.

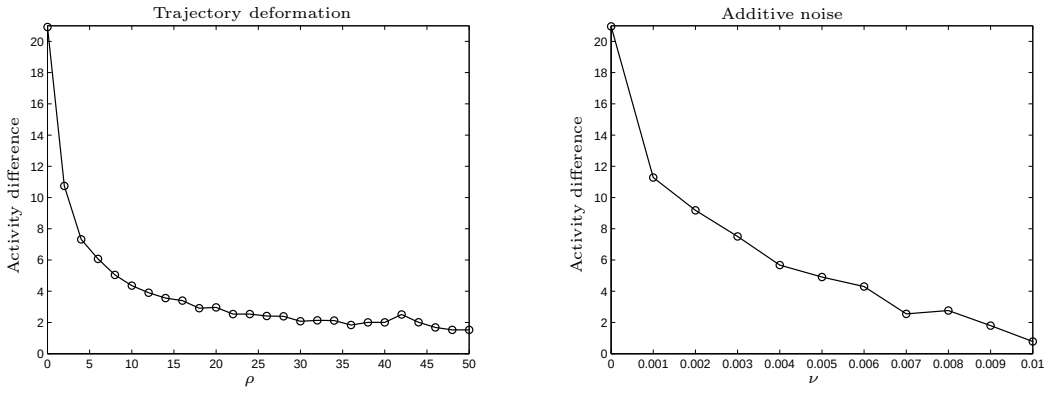
6.1.3 Generalization exploration

As we have presented, the ACNFT is able to classify simple, yet realistic, data. The ACNFT model, as a classifier, learns from a given instance (subject trial in our setup), and it should classify (discriminate) other instances (or subjects), otherwise the generalization capabilities of the model will be poor. To verify this property in our model, we design two variations of the sequences: trajectory deformations and additive noise. The first one corresponds to inter-subject and inter-trials variability, the second one is a standard noise model, that could correspond to some acquisition error (like computing motion in noisy video sequences), and evaluate the discrimination performance of the ACNFT model.

Trajectory deformation

Given a certain trajectory to be encoded by the ACNFT model, noted as $\vec{x}^i = (x^i, y^i)$, at each point in the trajectory (i), we consider a new position defined by $\vec{x}^i + N(\vec{0}, \rho)$, see Figure 6.7, where $N(\vec{0}, \rho)$ is a Gaussian noise function around 0. The perturbation is larger as ρ increases and several trials must be performed because of the variability of the random function, in this experiments 100 trials are considered. The training of our model was made with the original trajectories, then the modifications were introduced in the input, see Figure 6.7(a). The experiment was performed over the 3 sequences, then the mean variation among the population was computed: if the input is the "to clap" sequence, the population m_1 should have a larger activity than the other two, by consequence a positive difference. This difference was computed for each population, and the mean value was considered.

The results of the trajectory deformation can be seen in Figure 6.8(a). To compare the effect of both kinds of noise, we define an arbitrary discrimination level, as the noise level when the total activity of the right population is at least twice larger than for the other populations. In these experiments, this level is reached approximately at $\rho = 7$. This value is close to the width (3 that corresponds to $dr = 1$) of the modified trajectories + the size of the positive part of the



(a) (b)

FIGURE 6.8 – The mean minimal difference of the total activity between the input that corresponds to the input and the other populations, for the 3 actions and different noise level. For example, activity difference at 21 (no noise in (a) and (b)) corresponds to the mean of the diagonal (26.5) in Table 6.1 minus the mean of the values outside the diagonal (5.60) in the same table. (a) Trajectory noise. (b) Additive noise.

kernel w (4). We conclude that the tolerance to deformations in the trajectory, that could be associated to either inter-subject or trial variability in our model, is closely related to the size of the kernel (σ) in the ACNFT model + the width of the trajectory (dr).

Additive noise

The additive noise corresponds to adding a value between $\in [0, 1]$ (normalized image), with a Gaussian distribution to any point of the input, independently of the trajectory. The training trajectory is not deformed, but the final input receives the additive noise, see Figure 6.7(b). The results of the trajectory deformation can be seen in Figure 6.8, where it can be noticed that with a noise level of $\nu = 0.004$, the discrimination is no longer possible (using the same criteria as in the trajectory deformation). The noise amplitude, up to $\nu = 0.01$ or $(SignalAmplitude/NoiseAmplitude)^2 = 100$ or SNR [13], seems to considerably affect the ACNFT model. Probably, this is due to the linear combination of input and population activity, and because the mean activity of the population (associated to the discrimination) is performed over the complete space. We conclude that in noisy conditions, it is important to consider the ACNFT model together with denoising techniques, in order to reduce the neural activity into a set of local maxima, idea that we develop in the next section.

6.2 Feature extraction & discrimination

One of the questions we study in this thesis is the precise nature of the features in the brain such as to perform sequence discrimination using movement information in a distributed framework. The most straight-forward feature is to directly consider the raw optical flow, as some authors have proposed [EMVK09]. Other authors propose that local flow patterns may be the key feature to perform the discrimination [CG05]. The local flow patterns idea has strong

13. Signal to Noise Ratio

biological inspirations as this kind of operators has been reported in the MST area, and lesions to this area seem to reduce biological motion recognition rates [PM94].

In the previous section we evaluated our proposed classification mechanism. We showed its performance for the PL stimuli (few moving points) and discussed its properties. The proposed model can be considered as a metric to discriminate patterns, so that if we compare the same patterns using different features, we can evaluate the features using our ACNFT model. Of course, this analysis is limited to the kind of movement we are evaluating, in this case human motion, and starting from the idea that the metric (our ACNFT model) is related to the human performance to discriminate visual sequences.

In this section we compare two features computed from the "to clap", "to fight" and "to wave" sequences: the raw optical flow, and local pattern information. We describe now each feature, and then we present the discrimination results obtained by the ACNFT model using the raw optical flow and the local pattern information.

6.2.1 Raw Optical Flow

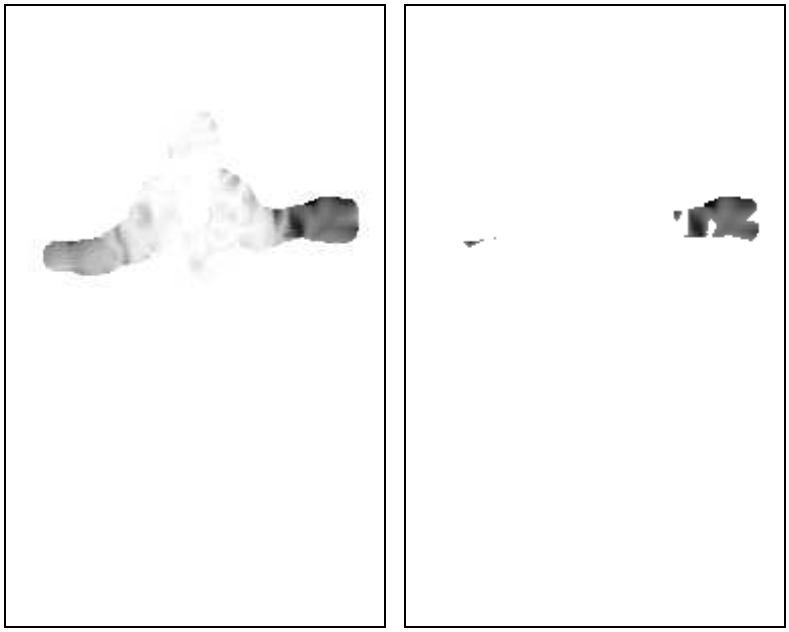
We have described how to perform the optical flow extraction in Chapter 1 and Chapter 3. Here we consider the multi-scale serial Lucas & Kanade algorithm, for the sake of simplicity. One of the characteristics of the optical flow is the presence of noise when detection is performed with real images. Extraction techniques, like Lucas & Kanade, tend to perform noisy detection or to produce artifacts, and they are also subject to the aperture problem, as we discussed in Chapter 2.

In the differential technique of Lukas & Kanade [Luc85], there is the idea of a small receptive field as found in the primary visual cortex, so in that sense it is still a local motion computation method. At each location (pixel) where the optical flow is computed, see Figure 6.9(a), we assign one unit, used as input by each 2D ACNFT population. However, we keep detection information only in a few places where we know that the detection has a good quality, using a threshold t_h (other locations are set to zero), i.e. mostly in spatial locations where corner-like features are present. This idea provides a way to filter the optical flow, see Figure 6.9(b). Empirically, this filtering applied to biological motion sequences highlights locations roughly associated with joints movements. To further simplify the optical flow we consider a fixed threshold for all sequences ($|v_{min}|$).

We sum up the raw optical flow extraction as,

1. Compute the partial derivatives I_x, I_y, I_t.
2. At each pixel where $det(\begin{bmatrix} W^2 * I_x^2 & W^2 * (I_x I_y) \\ W^2 * (I_x I_y) & W^2 * I_y^2 \end{bmatrix}) > t_h$ compute the optical flow as in Eq. 6.1.
3. Each pixel where $|v| < |v_{min}|$ is set to zero.
4. The value of $|v(\vec{x})|$ at each location delivers the input for the ACNFT model at $\vec{x}$.

$$\vec{v} = \begin{bmatrix} W^2 * I_x^2 & W^2 * (I_x I_y) \\ W^2 * (I_x I_y) & W^2 * I_y^2 \end{bmatrix}^{-1} \begin{pmatrix} W^2 * (I_x I_t) \\ W^2 * (I_y I_t) \end{pmatrix} \tag{6.1}$$



(a) (b)

FIGURE 6.9 – The raw optical flow. (a) Darker values represent higher values and areas with faster movement. (b) The same as (a) but keeping only the higher half of values.

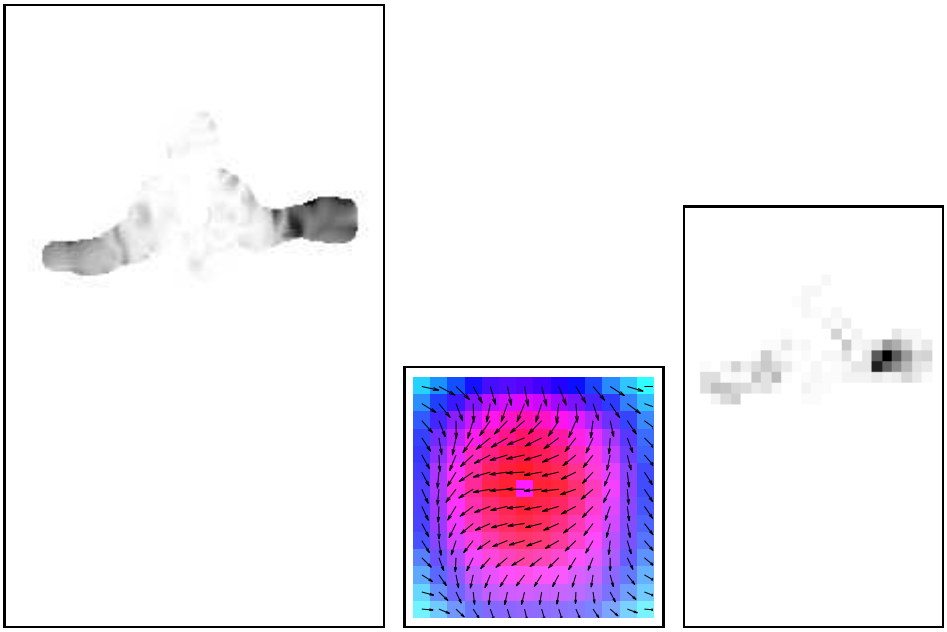
6.2.2 Local flow patterns

Local flow patterns have been described in the literature before, as mentioned in Chapter 4. Other than the anecdotic description of their receptive fields, the more precise nature of these pattern selective neurons is not well understood. However, some authors have proposed that it may be related to some statistical significance applied to the optical flow [CG05]. To check this, we consider the "to clap", "to fight" and "to wave" sequences, where we extract the optical flow, exactly as described in section 6.2.1, and look for the "best" local flow pattern to discriminate among the sequences (see Appendix C.1 for more details). The technique to select the best flow patterns was the PCA that we now describe shortly in our specific framework.

The Principal Component analysis (PCA) is a process to transform a given space from a set of correlated variables into a (reduced) set of uncorrelated variables [CM98]. If you consider a set of observations $\mathbf{x}_i$ each one of dimension n, the main idea is to retrieve a subset of c "components", such as to describe any observation $\mathbf{x}_i$ as in Eq. 6.2. For the optical flow, each observation is a small part of the visual field, in our case we consider a 14×14 pixels windows, so $n = 14 * 14 * 2 = 392$ (factor 2 appears because the velocity has two dimensions: direction and speed). We also consider a 50% overlap among observations, thus each 7 pixels we consider a visual patch of size 14×14, in agreement to observations that area MST presents a related degree of overlapping.

$$\mathbf{x}_i = \sum_c \alpha_i \mathbf{v}_i \tag{6.2}$$

In Eq. 6.2 α_i is a scalar value and $\mathbf{v}_i$ represents each one of the components. The $\mathbf{v}_i$ vectors will be fixed once the PCA is computed. When $c = n$, then the PCA preserves dimensionality, but we consider $c << n$ or the case where we reduce dimensionality. There are several ways to



(a) (b) (c)

FIGURE 6.10 – Absolute value of the optical flow. (a) Darker values represent higher values, thus areas of faster movement. (b) PCA Component, 14×14 pixels. (c) Projection of the OF over the PCA component in (b). The image is reduced by a factor of 7 because the projection is performed each 7 pixels (50% overlap).

obtain the $\mathbf{v}_i$ vectors, in this work we consider the Single Value Decomposition algorithm (SVD) detailed in Appendix C.1. Applied to optical flow sequences, with a receptive field of size 14×14 pixels, and a 50% overlap, selecting one component delivers a reduction by a factor of 7, because there is one value every 7 pixels. Thus if we consider up to seven components ($\mathbf{v}_i$) we are still reducing the space of variables.

The input to our ACNFT model are the $|\alpha_i|$ values at each location (less than pixels if we consider one component), and when several components are used we linearly combine the different $|\alpha_i|$. An example projection can be seen in Figure 6.10(c) using the component of Figure 6.10(b) to project on.

6.2.3 Features evaluation

We now present the discrimination result for the "to clap", "to fight" and "to wave" sequences, first using the raw optical flow signal and then the local optical flow patterns. The experiments where carried out over 64 frames from the animation of the sequences obtained by using the VICON© system, see Figures 6.2(d), 6.2(e) and 6.2(f). Each frame corresponds to a different t in the ACNFT model, and to compare discrimination values the mean activity $A(t)$ for each population was registered. The codification of sequences using the ACNFT was performed using one joint trajectory (left wrist), from the 2D projection from the VICON© system.

The results of the discrimination using the raw optical flow can be seen in Table 6.5. We can see that for the same input, the populations are able to distinguish among the different sequences (except for the "to clap" sequence), but the differences are rather weak (see by columns). However,

TABLE 6.5 – Confusion matrix for the classification using the raw optical flow. Populations are by rows, and different inputs by columns. $\sigma = 21$

	To clap	To fight	To wave
To clap (m_1)	25.50	42.66	23.42
To fight (m_2)	33.70	47.93	23.92
To wave (m_3)	25.58	40.28	24.68

TABLE 6.6 – Confusion matrix for the classification using the first PCA component. Populations are by rows, and different inputs by columns. $\sigma = 3$

	To clap	To fight	To wave
To clap (m_1)	18.06	5.48	6.94
To fight (m_2)	11.23	12.02	3.64
To wave (m_3)	3.91	3.51	10.45

if we consider several inputs for the same population (by rows), the discrimination is not possible. This is due to the fact that the ACNFT has a linear relation with the input (the raw optical flow) and different sequences have different amounts of "total movement", thus it is not possible to compare them this way. By consequence, one population using the raw optical flow as input cannot work, for example, as a binary classifier, because its activity will be determined by the total motion of the sequence. Compared with Table 6.1, the values in Table 6.5 are higher, specially outside the diagonal, because the sequences activate a large part of the neural field (both arms are moving), this is specially evident for the "to fight" sequence (that has a zone of high speed motion).

Next, we perform the discrimination using the local optical flow patterns, considering a single component. The results of this experiment can be seen in Table 6.6, where we can see that, as in the case of the raw optical flow, discrimination can be performed for the same input (by columns). The absolute values are lower than in the raw optical flow, due to the space reduction ratio that the selection of one PCA projection induces. An interesting difference with the raw optical flow is that the discrimination can be performed using a single population and using different inputs (by rows). This property is interesting because, even with sequences with different "total" movement, the PCA projection provides a "normalization" mechanism in terms of the total movement.

The following experiment considers the same conditions as in the first experiment with one PCA component, but now simultaneously using the first seven components, that provide 57% of the total variance, see Appendix C.2. The results of this experiment can be seen in Table 6.7. If we suppose that each component equally contributes to the discrimination, we should expect the values in the diagonal in Table 6.7 to be 7 times greater than in Table 6.6 because we now consider 7 components. However, the diagonal values are lower than expected, showing that the discrimination is not always improved considering more PCA components. Thus, some PCA components appear to contribute more in this discrimination task.

Local flow pattern over rotations

Our final experiment searches to verify if, beyond space reduction and motion normalization, the use of local optical flow patterns delivers some advantage in terms of rotation invariance. We know that PCA components are not invariant in the sense that, for example if the referential is rotated the component will also rotate, see Appendix C.1. Therefore, if they provide some

TABLE 6.7 – Confusion matrix for the classification using the firsts 7 PCA components. Populations are by rows, and different inputs by columns. $\sigma = 3$

	To clap	To fight	To wave
To clap (m_1)	105.20	44.44	37.22
To fight (m_2)	51.93	82.80	26.52
To wave (m_3)	48.08	27.58	66.50

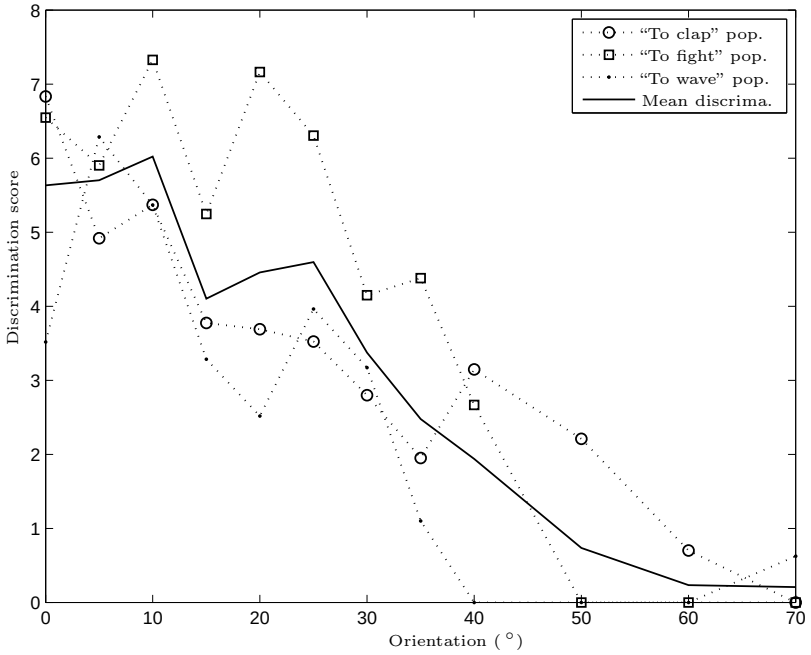


FIGURE 6.11 – Discrimination when one PCA component is analyzed, and the input is rotated considering the "to clap", "to fight" and "to wave" sequences. Also plot the mean discrimination.

degree of invariance it will be due to the spatial convergence, by consequence this invariance will be always partial. To verify this idea, we consider a single PCA component, as in Table 6.6 and we rotate only the input, populations being kept the same. Then, we define a discrimination score as the difference between the total activation of population that corresponds to the input and the closest different population (same as in 6.1.3), and plot the discrimination score for different rotations and populations, see Figure 6.11. It can be noticed that the score decreases, to be no longer significant at 30°. This contrasts with the 5° tolerance we obtain using directly the ACNFT (using the trajectories as input) in the first part of the chapter. We conclude that the convergence mechanism of local flow patterns seems to deliver a partial rotation invariance mechanism.

6.3 Discussion

In this chapter, we have evaluated our proposed ACNFT model for real movement sequences of PL stimuli and body animations. In all cases, our model shows a good capacity to discriminate the 3 movements we have considered: "to clap", "to fight" and "to wave". We also show that for the set of 3 sequences one trajectory is already sufficient to distinguish among them using either the PL or animation input. The general properties of our model were also explored showing extremely rapid classification, where only a few frames are necessary to discriminate among sequences. At the same time, we verify that tailoring certain units with a specific kernel (the asymmetric kernel) makes our model very sensitive to rotations. We also show that as more (or longer) trajectories are considered, the discrimination rate increases linearly. With respect to the generalization properties of the ACNFT model, we determine that it can handle variations (associated with inter subject/trial variations) up to the size associated to the trajectory width plus the positive part of the kernel ($\sigma/4$). When additive noise is added to the input (associated with noisy movement detection), we show that the ACNFT model requires a good SNR ($>> 1$) to correctly discriminate sequences. We also performed an exploratory work with a face movements database, to illustrate that the ACNFT could be applied to different kinds of spatiotemporal sequences not necessarily body motion, see Appendix D.

To summarize the first part of the chapter, we show that the distributed architecture proposed for the discrimination of spatiotemporal sequences is effective and remains possible to use with real sequences. At the same time, we reproduce a set of observed properties of the human biological motion recognition like rapid categorization, and partial invariance to rotations. We also notice that our model cannot explain random PL or ISI stimuli because a disruption in the local structures (the trajectories) interferes with the discrimination (this is not the case for humans). We argue that this problem may arise because our model is at the best incomplete, and does not include any pre-motor or form information to make it more robust to this kind of perturbation.

The second part of this chapter compares the discrimination rate achieved by our model either using the raw optical flow or local optical flow operators directly (associated with detectors of "discontinuity"). Our results show that the two kinds of inputs can be used for discrimination in our database, but better results (higher discrimination rates) can be achieved with the local optical flow operators. This result is interesting because optical flow operators provide not only a better discrimination rate (make the system more robust), but also a reduction in the dimensionality of the problem (the receptive fields comprehend a 14×14 pixel window) and a normalization mechanism. At the same time, we verify that space reduction increases the invariance to rotations, making the system partially invariant to rotations (up to $30\,^{\circ}$). This result may give statistical support to the evidence that MST injuries produce a reduction in the discrimination robustness in biological motion recognition.

Local optical flow patterns in conjunction with our ACNFT model deliver a functional distributed mechanism for the classification of temporal sequences and, specifically, for human motion sequences. Altogether, this distributed architecture exhibits rapid discrimination, incremental discrimination and partial rotation invariance. These results highlight the importance of local motion signals to describe visual sequences. Local motion information, combined with an implicit body model (the trajectories of movement), provides a hypothesis about how the recognition of biological motion may be performed in the brain: by sets of local descriptors following an implicit body model.

Part III

Conclusions

Conclusions

In this work, we have studied the computation of motion from a distributed perspective, comparing techniques in computer vision and current knowledge in biology. In the first part we focus on two problems of the local estimation of motion: the aperture problem and the speed sampling problem. For these problems we show that computing movement locally is not sufficient to obtain precise movement detection: integration must be performed within the population of local motion detectors. The integration can be performed through the dynamics of the populations of units (neurons) to find the correct motion estimation. The second part of this thesis explores the question of how to encode and retrieve complex visual patterns (like biological motion) in a distributed framework taking inspiration from brain processing of visual information. We show that the spatial structure of the visual sequence allows for recognition. Moreover, this spatial structure can be encoded as a set of (local) spatiotemporal trajectories in a proposed model (ACNFT) of visual sequences discrimination through the dynamics of a population of units.

In the second chapter, we develop one aspect of the location estimation of movement related to its orientation, the aperture problem. In the context of distributed computation and using a recursive mechanism that propagates (over space and time) coherent answers among neurons, we show that providing a semantic to what each unit processes (movement direction) helps to combine information by including inhibition between opposite orientations. More precisely, we have illustrated how in some cases this modification can make the process more rapid, but overall more robust when the initial motion estimations are noisy. The inhibition among local movement detectors can also be seen as a functional interpretation of the receptive field structure found in MT neurons, where the interaction is excitatory among similar orientations but seems to be inhibitory for opposite movement orientations .

Chapter 3 explores the multi-scale detection of speed, where we establish the link among relative error and discrimination of speed, metrics from computer vision and experimental psychophysics respectively. Once the link is establish, we propose that a population weighted average read-out with a log-scale sampling is sufficient to achieve flat speed discrimination curves as observed in humans. Moreover, the precise relative error can be controlled by the degree of overlapping among scales. When compared with computer vision serial algorithms for multi-scale estimation of speed, this kind of scheme avoids error accumulation by evaluating at all scales simultaneously. The simultaneous computations of speed at several scales suggest a scheme to combine multi-scale estimation for any motion detection algorithm in computer vision, where a certain degree of homogeneous relative error should be attained. At the same time, our work suggests a logarithmic distribution for speed sensitive neurons that has already been suggested by recent recordings of multiple neurons in primates.

In the second part (Chapters 4, 5 and 6) of this thesis, we study extensive experimental results about how the encoding and recognition of complex visual patterns are performed in the brain. As a summary, we conclude that local features (such as motion) are crucial to encode complex visual patterns. The precise detection is not as important as the spatiotemporal structure of the pattern, which seems to be critical. We find support for this argument, first in the psychophysics

bibliography, where it has been shown that local motion estimation can be disrupted, but complex patterns can still be recognized by human subjects. We also support this argument based on the experimental evidence that local estimation of motion requires computing time. This time is longer (or comparable) with higher cognitive tasks such as the recognition of biological motion. Thus, the precise movement orientation/amplitude does not seem the most important feature that the brain requires to encode complex patterns. Rather, its spatial structure appears essential.

After the introduction in Chapter 4 of the idea of local features to encode complex visual patterns, we conclude that the spatial structure of the movement seems to be the determinant factor to encode complex visual patterns, as shown by experiences such as the PL stimuli. Following this hypothesis and making the assumption that spatiotemporal trajectories are separable, we build a neural field model in Chapter 5 able to encode any set of local spatiotemporal trajectories. The main contribution of this model is to show, in the distributed framework of neural fields, that asymmetry in the connectivity between neurons is sufficient to encode local trajectories (orientation and speed). The former idea was demonstrated analytically and illustrated by computational simulations. In particular, we show a simple analytical relation to set the parameters of the populations of units to encode any given set of spatiotemporal trajectories. Simulations were carried out to show that the encoding of trajectories was effective for simple sequences. The model extends the existing 1D analysis of asymmetric neural fields to 2D and precises the dependency between the network connectivity (asymmetry) and the input speed, exploring the dynamics that such a neural population may have.

In the last chapter, we evaluate our model with real movements using single and two joints trajectories showing that both orientation and speed can be encoded by the proposed mechanism. Additionally, we show that several of the properties of human pattern recognition (in the case of biological motion) can be retrieved in our model, such as: rapid discrimination, tolerance to time-warping, incremental discrimination and view-dependent recognition. We conclude that our hypothesis of a distributed representation of visual patterns by representing their spatial structure can, at least, retrieve some of the observed properties of human capabilities for this task. Finally, we evaluate the capacity of our model to encode complex visual patterns using either the raw optical flow or intermediate optical flow operators directly as input. The two local features are noisy and present intrinsic problems such as the aperture problem. In this respect, we conclude that, even though sequences can be encoded by using the raw optical flow, the use of optical flow patterns presents numerous advantages, such as increasing the invariance to rotations and normalizing the sequences. To conclude, our model presents a functional architecture from a computer vision perspective. It also presents a hypothesis of how visual patterns may be encoded in the brain: by means of populations of neurons encoding the spatial structure of movement. This is an alternative to the hypothesis of "snapshots" neurons in area STS/FBA where each instant of a sequence could be encoded. At the same time, we show that the functional role of the local flow operators in area MST may be interpreted as a denoising and normalization stage to empower the discrimination system on top of it. The raw optical flow can still be used as input, although it induces a discrimination that could be considerably more sensitive to noise (as has been observed when area MST presents damage).

The main contributions of this thesis are as follows:

Aperture problem: This thesis presents a model of motion detection in area V1/MT. This model, composed of a distributed population of units, successfully delivers a coherent answer, i.e. it solves the aperture problem. It works even in the presence of noise by using populations of local velocity estimators combined in a recursive architecture with excitatory and inhibitory connections.

Multiscale speed detection: A metric is proposed in Chapter 3 to compare speed discrimination in experimental psychophysics and computer vision experiments. We also propose an algorithm to achieve speed discrimination comparable to experimental psychophysical results using any optical flow technique. This algorithm determines the distribution of the local speed detectors, allowing broader detection larger than the range of any particular detector with a uniform discrimination quality in this range.

Visual sequences coding: In the second part of this thesis, we show how a distributed model to perform pattern recognition is able to encode complex visual sequences by using local features, where neuronal populations represent the structure of the movement (like a body model in the case of biological motion).

Brain mechanism: At the same time, this latter model presents a hypothesis about how the visual pattern recognition in the brain may be performed, not by saving the moving/static templates (or snapshots), but by keeping track of the dynamics of local features.

Local optical flow patterns: We also show that local optical flow patterns in our sequence discrimination model, contribute to improving the invariance to rotations, normalization and in general, to the discrimination of visual sequences. This idea proposes an interpretation of the functional role of area MST in the encoding and retrieval of visual sequences.

Distributed computation: Taken together, the ideas presented in this thesis show how a difficult task such as movement detection and coding/recognition of visual sequences can be performed in a connectionist framework, taking inspiration from the brain. Our models support the idea that populations of units with a small receptive field can perform complex operations by means of combining their information. At the same time, the models presented provide functional interpretations of experimental evidence from biology and interesting insights for computer vision, such as the aforementioned recursive architecture, the multi-scale architecture for movement detection and the encoding of visual sequences by mean of local features.

Future work & perspectives

The main perspectives of this thesis work are twofold. The first is related to the application of the presented ideas to standard sequences databases in computer vision to compare the classification performance of the proposed model to other available methods. Second, the exploration in partnership with biologists of the precise brain mechanisms responsible for some aspects of detecting local movement or encoding complex visual sequences, especially asymmetry.

The application of our model to standard human sequences (like the KTH database) represents the most short-term perspective of our work. Here, the main interest is to evaluate the relative performance with respect to techniques from computer vision, either based on local or global features. The main issue with this kind of database is how to break down the visual sequences into sets of local trajectories. It is very likely that this task needs to be performed manually. One alternative to avoid this very time-consuming task is to design an internal body

representation to help out in the learning stage. However, the application of our model to video sequences could provide a very interesting perspective, even if the learning stage is performed manually.

One of the extensions to be considered by our model of motion processing between areas V1 and MT for the aperture problem is the integration of disparity information (many of the neurons in MT are sensitive to disparity). Such a model could greatly benefit from collaboration with biologists to analyze the dynamics in MT related to disparity. The verification of the spatial distribution of speed selective neurons that we propose in this thesis (log distributed) could be studied in collaboration with biologists. As well, the theoretical results of Chapter 5 could be developed in cooperation with neurobiologists to verify the properties of asymmetric activity propagation in cortical tissues. More precisely, our theoretical results suggest how pulse stimuli may behave when there are both excitatory and inhibitory lateral connections. All these perspectives represent medium and long-term extensions to my work, where the collaboration with biologists will be paramount to understand how the brain processes visual information.

Another aspect to be developed as an extension to this thesis is the integration of top-down and learning mechanisms in the distributed model of visual sequence discrimination proposed in Chapter 5. The experimental evidence that local information is not absolutely necessary to discriminate is very intriguing. This idea challenges several computational models, including our model introduced in Chapter 5. In the work of Thornton et al [TI98], inter-stimuli blank frames (ISI) were inserted in the PL stimuli to avoid the activation of neurons with short integration time (such as in MT and MST), showing that discrimination is still possible. It implies support for top-down influences. The same is suggested by the joint-displaced stimuli experiment, where not the joints, but rather the middle-bone positions were marked [BP94], or even the random PL stimuli experiment [BP94]. In fact, feed-forward models (like ours) where the receptive field size increases along the processing path cannot explain these results. One hypothesis that has been formulated to address this problem is the existence of an internal "simulator" able to estimate the inverse kinematics and to anticipate human movements. Such anticipation could allow the brain to avoid local perturbations by using this top-down signal as a reinforcement of coherent movement. A good candidate to support this hypothesis is provided by mirror neurons [RFG01] reported to be active whenever our vision detects someone moving. Some clues in this direction have already been given in [JS04] where the execution of a motor task shows a significant influence over the visual perception of the same motor task or other tasks, i.e. if the observer is walking his ability to distinguish other walking subjects will drop. These aspects are interesting to be included as an extension to the visual sequence discrimination model we present, for example by integrating for instance a bio-mechanical body model to dynamically infer the local movement patterns that characterize each visual sequence.

A

Asymmetric neural fields

A.1 1D ACNFT

A.1.1 CNFT transformation

Let us consider Eq. 5.1, following the idea by Xie et al [XG02] to change the non-linearity from the integral to the integral plus the input to simplify further analysis. We can call the input $I(x,t) = E(x,t) + h$ and write it as $I(x,t) = \tau\frac{\partial \tilde{I}(x,t)}{\partial t} + \tilde{I}(x,t)$, using this we can rewrite Eq. 5.1 as,

$$\frac{\tau\partial}{\partial t}\left(u(x,t) - \tilde{I}(x,t)\right) + u(x,t) - \tilde{I}(x,t) = \int_\Omega w(|x'-x|)f[u(x',t)]dx' \tag{A.1}$$

Assuming that the activity u can be written as $u(x,t) - \tilde{I}(x,t) = \int_\Omega w(|x'-x|)m(x',t)$, and as w can be in general any function, by consequence the integral can be removed in each of the three terms, to express Eq. A.1 as,

$$\tau\frac{\partial m(x,t)}{\partial t} + m(x,t) = f\left[\int_\Omega w(x'-x)m(x',t)dx' + \tilde{I}(x,t)\right] \tag{A.2}$$

This kind of model has already been proposed [HS98], the main difference between the classical neural field equation in Eq. 5.1 and Eq A.2 is the relation between the variation of the neural activity and the external input. In Eq. A.2 there is no variations at saturation levels. The next simplification we use is to avoid border effects by integrating in a ring, thus integrating over a circular domain,

$$\frac{\partial m(\theta,t)}{\partial t} + \tau m(\theta,t) = f\left[\int_{-\pi}^{\pi} w(\theta'-\theta)m(\theta',t)d\theta' + I(\theta,t)\right] \tag{A.3}$$

A.1.2 Analytical expression of $r_0(t)$

This first analysis has been performed elsewhere (see [XG02, HS98]), and we only make explicit and clarify some of the steps. Given Eq. 5.1, we first look for a closed form for the mean activity (first Fourier component) $r_0(t)$,

$$r_0(t) = \int_{-\pi}^{\pi} m(k,t)(2\pi)^{-1}dk \tag{A.4}$$

To achieve this we use the same input as in [XG02, HS98], see Eq. A.5, at $t = 0$ this represents a bump around $k = 0$,

$$I(k,t) = C\left[1 - \epsilon + \epsilon\cos(k - vt)\right] - T \tag{A.5}$$

shifting towards the right at speed v, where the parameters C, ϵ and T control the ratio between maximal $(C - T)$ and minimal $(C - T - 2\epsilon)$ activity. We also need the dynamics of the second Fourier component $r_1(t)$ defined in Eq. A.6, because $r_0(t)$ depends on it, as we will see later on.

$$r_1(t) = \int_{-\pi}^{\pi} m(k',t)(2\pi)^{-1}e^{i(k'-\psi(t))}dk' \tag{A.6}$$

In Eq. A.6 $\psi(t)$ is an unknown function of time, that makes $r_1(t)$ a real and positive number, in other words for a single bump solution, this is the peak of the bump position (or phase in the complex plane). Introducing Eq. A.5 and Eq. 5.4 into Eq. 5.1 and identifying r_0 as defined in Eq. A.4 we can rewrite Eq. 5.1 as,

$$\begin{aligned}\tau\frac{\partial m(k,t)}{\partial t} + m(k,t) = \Bigg[& J_0 r_0(t) + C(1-\epsilon) - T + \\ & C\epsilon\cos(k - vt) + J_1\int_{-\pi}^{\pi} m(k,t)(2\pi)^{-1}\cos(k' - k + \beta)dk'\Bigg]^+\end{aligned} \tag{A.7}$$

here the point is to simplify as much as possible any dependency in the space, and to impose a single bump solution. To do this, we can rewrite the right-side of Eq. A.7 in the complex plane to simplify calculations (taking only the real part) and then using Eq. A.6,

$$\begin{aligned}
&J_0 r_0 + C(1-\epsilon) - T + C\epsilon e^{-i(k-vt)} + J_1\int_{-\pi}^{\pi} m(k,t)(2\pi)^{-1}e^{i(k'-k-\beta)}dk' \\
&J_0 r_0 + C(1-\epsilon) - T + C\epsilon e^{-i(k-vt)} + J_1\int_{-\pi}^{\pi} m(k,t)(2\pi)^{-1}e^{i(k'-k-\beta+\psi-\psi)}dk' \\
&J_0 r_0 + C(1-\epsilon) - T + C\epsilon e^{-i(k-vt)} + J_1\int_{-\pi}^{\pi} m(k,t)(2\pi)^{-1}e^{i(k'-\psi)}e^{i(-k-\beta+\psi)}dk' \\
&J_0 r_0 + C(1-\epsilon) - T + C\epsilon e^{-i(k-vt)} + J_1 r_1(t)e^{i(-k-\beta+\psi)}
\end{aligned} \tag{A.8}$$

To ensure a bump solution, we assume a single cosine shape as total input (all the terms inside $[\ ]^+$) introducing another variable: $\phi(t)$, which represents the spatial position of the total input center in Eq. A.8 (not in m as $\psi(t)$). Taking only the real part of Eq. A.8, and introducing ϕ we obtain,

$$\begin{aligned}I_0(t) + &\cos(\phi - k)(J_1 r_1\cos(\psi - \beta - \phi) + C\epsilon\cos(vt - \phi)) + \\ &\sin(\phi - k)(J_1 r_1\sin(\psi - \beta - \phi) + C\epsilon\sin(vt - \phi))\end{aligned} \tag{A.9}$$

where $I_0(t) = J_0 r_0(t) + C(1-\epsilon) - T$. The single bump shape for the solution translates then into imposing Eq. A.10.

$$J_1 r_1 \sin(\psi - \beta - \phi) + C\epsilon \sin(vt - \phi) = 0 \tag{A.10}$$

Using Eq. A.9, the condition in Eq. A.10 and the auxiliary variable $I_1(t) = J_1 r_1 \cos(\psi - \beta - \phi) + C\epsilon \cos(vt - \phi)$, we can rewrite Eq. 5.1 as:

$$\tau \frac{\partial m(k,t)}{\partial t} + m(k,t) = [I_0(t) + \cos(k - \phi(t)) I_1(t)]^+ \tag{A.11}$$

Until now, we have imposed a shape for the solution (single bump) and thus we have obtained a restriction (Eq. A.10). Rewriting the system, we have derived Eq. A.11. The next step is to deal with the non-linearity $[\quad]^+$ and to show the existence and the stability of the solution, *i.e.* $dr_0(t)/dt = 0$ and $dr_1(t)/dt = 0$. In order to simplify the non-linearity we use variable $k_c(t)$ (the unknown width of the total input bump). When $k_c = k - \phi$, the right side of Eq. A.11 is zero and we obtain $I_0(t) = -I_1(t)\cos(k_c)$, this allows us to write,

$$\tau \frac{\partial m(k,t)}{\partial t} + m(k,t) = [I_1(t)\,(\cos(k - \phi(t)) - \cos(k_c))]^+ \tag{A.12}$$

Before using $dr_0(t)/dt = 0$ and $dr_1(t)/dt = 0$, the last step is to transform Eq. A.12 into the Fourier domain by integrating each term in Eq. A.12 with $\int_{-\pi}^{\pi}(2\pi)^{-1}dk$ for the first Fourier component, and with $\int_{-\pi}^{\pi}(2\pi)^{-1}e^{ik}dk$ for the second component. The second component derives into two equations since it is complex, describing the dynamics of the system with a total of 3 coupled equations in the Fourier domain, see Eqs. A.13, A.14 and A.15.

$$\tau \frac{\partial r_0(t)}{\partial t} + r_0(t) = I_1(t) f_0(k_c) \tag{A.13}$$

$$\tau \frac{\partial r_1(t)}{\partial t} + r_1(t) = I_1(t) f_1(k_c)\cos(\phi - \psi) \tag{A.14}$$

$$\tau r_1(t) \frac{\partial \psi(t)}{\partial t} = I_1(t) f_1(k_c)\sin(\phi - \psi) \tag{A.15}$$

f_0 and f_1 are increasing functions of k_c defined as in [XG02, HS98]. The system is coupled because $I_1(t)$ depends on $r_1(t)$. Now, we can impose $dr_0(t)/dt = 0$, $dr_1(t)/dt = 0$ and $d\psi(t)/dt = v$ (input and solution for $m(k,t)$ move at the same speed), to finally obtain the closed form of r_0 and r_1,

$$r_0(k_c) = f_0(k_c)\frac{S}{-J_0 f_0(k_c) - \cos(k_c)} \tag{A.16}$$

$$r_1(k_c) = f_1(k_c)\frac{S}{-J_0 f_0(k_c) - \cos(k_c)}\cos(\Delta) \tag{A.17}$$

where $S = (C(1-\epsilon) - T)$ and $\Delta = \arctan(\tau v)$. This solution exists (and if it exists we know it is stable) if the condition in Eq. A.10 can be achieved. To check this we inject Eqs. A.16 and A.17 into Eq. A.10, obtaining:

$$S' = \frac{J_0 f_0(k_c) + \cos(k_c)}{\sqrt{J_1^2 f_1^2 \cos(\Delta)^2 - 2J_1 f_1 \cos(\Delta)\cos(\Delta+\beta) + 1}} \tag{A.18}$$

where $S' = (1-(C\epsilon/C-T)^{-1})^2$. Then the system has a solution, which is stable, only if Eq. A.18 can be verified for a given set of parameters. This verification cannot be performed analytically (we do not know k_c), but numerically changing v and β and fixing the other parameters there is a range of parameters where Eq. A.18 can be verified.

A.1.3 Optimization of $v(\beta)$

The analysis detailed now is original from this work and it differs from [HS98] in the asymmetry and in the results we obtain and the limits we verify from [XG02].

Once the shape of $r_0(t)$ has been determined, see Appendix A.1.2 for details, we want to optimize it as a function of the input speed v. In other words, we want $\beta(v)$ to maximize r_0 for stable solutions, *i.e.* that verify Eq. A.18.

First, it can be noticed from Eq. A.19 that maximizing $r_0(k_c)$ is equivalent to minimizing k_c in Eq. A.16 as f_0 is an increasing function of k_c, $J_0 < 0$ and $S > 0$.

$$r_0(k_c) = \frac{S}{-J_0 - \cos(k_c)/f_0(k_c)} \tag{A.19}$$

In the other hand, Eq. A.18 can be derived in terms of v and setting the extrema at zero *i.e.* $\partial k_c/\partial v|_{v=v_m} = 0$, we can obtain the minimum k_c in terms of v, or v_m. If we apply the operator $\frac{\partial}{\partial v}$ in Eq. A.18, we obtain two solutions for v_m:

$$v_m = \frac{J_1 f_1(k_c)\tau - 2\tau\cos(\beta) \pm \sqrt{J_1^2 f_1^2(k_c)\tau^2 - 4\tau^2 J_1 f_1(k_c)\cos(\beta) + 4\tau^2}}{2\tau^2 \sin(\beta)} \tag{A.20}$$

To choose between the two solutions, we verify that at the limit $J_1 f_1 \to \frac{1}{cos(k_c)}$, the solution must verify the only admissible solution $v = \tan(\beta)/\tau$ to the system, see details on how to obtain this limit in [XG02]. This makes us choose the "$-$" solution. It can be noticed that as the optimization starts with a stable solutions, the solution we found belongs to this regime. This remark is important, as this kind of system usually gets unstable as the input speed is more distant from the intrinsic speed of the system (usually as lurching waves, see [FB05]). To determine the range of operation (where activity is stable), and to obtain the small perturbations analysis a linearization of the system must be performed to obtain its eigenvalues, for more details check the chapter on small perturbations in [GP01], and the application to this case in [FB05, HS98, XG02].

A.2 2D ACNFT

A.2.1 Analytical expression of $r_0(t)$, *2D*

The following *2D* analysis has not been performed elsewhere to our knowledge. The idea is to follow the *1D* idea, and to find an expression for the first Fourier component, or the mean total activity and 3 other Fourier components needed to express the dynamics of the system. The first step is to assume an input function,

$$g(I(\vec{x},t)) = E(\vec{x},t) = D^2\left[1+\cos(y)\right]\left[1+\cos(x-vt)\right] + C \tag{A.21}$$

where D and C are parameters controlling the bump highest activity and the spontaneous activity level. If $x \in [-\pi,\pi]$ and $y \in [-\pi,\pi]$, then $E(\vec{x},t)$ defines a single bump moving from left to right ($v > 0$). For the kernel function we use the same equation as in Eq. 5.11,

$$w(\vec{x},\vec{x}') = A\left[1+\cos(y'-y)\right]\left[J_0 + J_1\cos\left(x'-x+\beta\right)\right] \tag{A.22}$$

where the asymmetry is only in the x-axis, therefore we are considering only sequences moving on that axis. If $y \in [-\pi,\pi]$, then w is a decreasing function of y. As in *1D* we start by writing the dynamics of $m(\vec{x},t)$ in terms of quantities that are space independent (Fourier components). We require first the definition of $r_0(t)$ in *2D*, see Eq. A.23,

$$r_0(t) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} m(\vec{x},t)d\vec{x} \tag{A.23}$$

but we will also require higher Fourier terms:

$$r_1(t) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} m(\vec{x},t)e^{i(x-\psi(t))}(2\pi)^{-2}d\vec{x} \tag{A.24}$$

$$r_2(t) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} m(\vec{x},t)e^{i(y-\Omega(t))}(2\pi)^{-2}d\vec{x} \tag{A.25}$$

$$r_3(t) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} m(\vec{x},t)e^{i(x+y-\lambda_1(t))}(2\pi)^{-2}d\vec{x} \tag{A.26}$$

$$r_4(t) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} m(\vec{x},t)e^{i(x-y-\lambda_2(t))}(2\pi)^{-2}d\vec{x} \tag{A.27}$$

where the unknown functions of time $\psi(t)$, $\Omega(t)$, $\lambda_1(t)$ and $\lambda_2(t)$ are such that they make each quantity real and positive. Before using the different Fourier components, we express the dynamics of m using Eq. A.22,

$$\begin{aligned}\tau\frac{\partial m(\vec{x},t)}{\partial t} + m(\vec{x},t) = \Bigg[E(\vec{x},t) + AJ_0r_0(t) + \frac{AJ_0}{4\pi^2}\int\int cos(y'-y)m(\vec{x}',t)d\vec{x}' + \\ \frac{AJ_1}{4\pi^2}\int\int\cos(x'-x+\beta)m(\vec{x}',t)d\vec{x}' + \frac{AJ_1}{4\pi^2}\int\int cos(y'-y) \\ \cos(x-x'+\beta)m(\vec{x}',t)d\vec{x}'\Bigg]^+\end{aligned} \tag{A.28}$$

All the right side of Eq. A.28 is the total input, and the $[\quad]^+$ operator must be applied to it. Now, we rewrite the total input in the complex plane to simplify,

$$\begin{aligned}&= \Bigg[E(\vec{x},t) + AJ_0r_0(t) + \frac{AJ_0e^{i(y-\Omega)}}{4\pi^2}\int\int e^{-i(y'-\Omega)}m(\vec{x}',t)\vec{x}' + \frac{AJ_1e^{-i(x-\beta-\psi)}}{4\pi^2} \\ &\int\int e^{i(x'-\psi)}m(\vec{x}',t)\vec{x}' + \frac{AJ_1}{2}\Bigg\{e^{i(x+y-\lambda_1+\beta)}\int\int\frac{e^{i(x'+y'-\lambda_1)}}{4\pi^2}m(\vec{x}',t)\vec{x}' + \\ &e^{i(x-y-\lambda_2+\beta)}\int\int\frac{e^{i(x'-y'-\lambda_2)}}{4\pi^2}m(\vec{x}',t)\vec{x}'\Bigg\}\Bigg]^+\end{aligned} \tag{A.29}$$

where we omit the limits of the integrals as they are all either π or $-\pi$. Introducing now Eqs. A.24, A.25, A.26 and A.27 we can write more compactly Eq. A.29 as,

$$= \Bigg[E(\vec{x},t) + AJ_0 r_0(t) + AJ_0 e^{-i(y-\Omega)} r_2(t) + AJ_1 e^{-i(x-\beta-\psi)} r_1(t) + \frac{AJ_1}{2} \Big\{ e^{i(x+y-\lambda_1+\beta)} r_3(t) + e^{i(x-y-\lambda_2+\beta)} r_4(t) \Big\} \Bigg]^+ \tag{A.30}$$

In the complex plane, we will impose a total input with the shape of a single bump moving along the x direction. This implies that the modes associated to $x+y$ and $x-y$ must have the same coefficients, *i.e.* $r_3 = r_4$ and $\lambda = \lambda_1 = \lambda_2$. Using this and Eq. A.21, the total input can be written as:

$$= \Big[I_0(t) + e^{i(x-\phi)} I_1(t) + e^{-i(y-\Omega)} I_2(t) + I_3(t) e^{-i\alpha} (e^{i(x+y)} + e^{i(x+y)}) \Big]^+ \tag{A.31}$$

where $I_0(t) = AJ_0 r_0(t) + C + D^2$, $I_1(t) = AJ_1 e^{i(\beta-\psi+\phi)} r_1(t) + D^2 e^{i(\phi-vt)}$, $AJ_0 e^{-i\Omega} r_2(t) + D^2$ and $I_3(t) = (AJ_1 e^{i(\beta-\lambda+\alpha)} r_3(t) + D^2 e^{i(\alpha-vt)})/2$. In Eq. A.31 we have also introduced two variables $\phi(t)$ and $\alpha(t)$, that are used later to ensure a single bump. Eq. A.31 can be simplified into Eq. A.32 using the identity $e^{i(x+y)} + e^{i(x-y)} = 2\cos(y)e^{ix}$.

$$= \Big[I_0(t) + e^{i(x-\phi)} I_1(t) + e^{-i(y-\Omega)} I_2(t) + I_3(t) 2\cos(y) e^{i(x-\alpha)} \Big]^+ \tag{A.32}$$

As in the *1D* case, we want a single bump for the total input and we impose $I_1(t)$ to be real and to ensure that we introduce $\phi(t)$ similarly for $I_3(t)$ and $\alpha(t)$. Also in *2D*, we want $\Omega(t) = 0$ as the total input should remain static in the y-axis. Furthermore, the total input expressed in Eq. A.32 requires $\psi(t) = \lambda(t)$ and $\phi(t) = \alpha(t)$ to be a single bump, which implies $r_1(t) = r_3(t)$, otherwise two bumps could exist. Using these assumptions, and taking back Eq. A.32 into the Real domain, we obtain,

$$= [I_0(t) + I_1(t)\cos(x-\phi) + I_2(t)\cos(y) + I_1(t)\cos(y)\cos(x-\phi)]^+ \tag{A.33}$$

There is one aspect of Eq. A.33 that requires to be addressed; to have a symmetric single bump we must have $I_1(t) = I_2(t)$ or, using both definitions, $A(J_0 r_2 - J_1 r_1 \cos(\beta+\psi+\phi)) = D^2(\cos(\phi - vt) - 1)$, in other words $r_1(t) \propto r_2(t)$. This property will help to obtain $r_0(t)$. Using this assumption Eq. A.33 can be simplified into Eq. A.34

$$= [I_0(t) + I_1(t)\cos(x-\phi) + I_1(t)\cos(y) + I_1(t)\cos(y)\cos(x-\phi)]^+ \tag{A.34}$$

The total input is a single bump in *2D* of width x_c and height y_c, using the symmetry of the total input we get $x_c = y_c$. At point (x_c, x_c) we know that the total input is zero or $I_0 = -I_1(2\cos(x_c) + \cos(x_c)^2)$, allowing us to write:

$$= I_1(t)\left[\cos(x-\phi) + \cos(y) + \cos(y)\cos(x-\phi) - 2\cos(x_c) - \cos(x_c)\cos(x_c)\right]^+ \tag{A.35}$$

Eq. A.35, implies that we can describe the dynamics of m in *2D*, using two Fourier components: $r_0(t)$ and $r_1(t)$ as in the *1D* case, and integrating Eq. A.35 by $\int\int d\vec{x}'/(2\pi)^2$ and $\int\int e^{ix} d\vec{x}'/(2\pi)^2$:

$$\tau\frac{\partial r_0(t)}{\partial t} + r_0(t) = I_1(t)g_0(x_c) \tag{A.36}$$

$$\tau\frac{\partial r_1(t)}{\partial t} + r_1(t) = I_1(t)g_1(x_c)\cos(\phi-\psi) \tag{A.37}$$

$$\tau r_1(t)\frac{\partial \psi(t)}{\partial t} = I_1(t)g_1(x_c)\sin(\phi-\psi) \tag{A.38}$$

where the only differences with the *1D* case are functions g_0 and g_1 that we define in Eq. A.39 and Eq. A.40.

$$g_0(x_c) = \frac{1}{\pi^2}\int_0^{x_c}\int_0^{l(x_c)}(\cos(x)+\cos(y)+\cos(y)\cos(x)-K(x_c))d\vec{x} \tag{A.39}$$

$$g_1(x_c) = \frac{1}{\pi^2}\int_0^{x_c}\int_0^{l(x_c)}(\cos(x)+\cos(y)+\cos(y)\cos(x)-K(x_c))\cos(x)d\vec{x} \tag{A.40}$$

using $K(x_c) = 2\cos(x_c)+\cos(x_c)^2$. The function to be integrated in Eq. A.39 represents a bump, with the same form in the four quadrants (symmetry). More precisely $l(x_c) = \arccos[(K(x_c) - \cos(y))/(1+\cos(y)]$.

The set of equations A.36, A.37 and A.38 describes the dynamics of $m(\vec{x},t)$ and as we want a stable activity, we impose $dr_0/dt = 0$, $dr_1/dt = 0$ and $d\psi(t)/dt = v$ and use $I_0 = -I_1(2\cos(x_c)+\cos(x_c)\cos(x_c))$ to obtain,

$$r_0(x_c) = \frac{(C+D^2)}{-AJ_0 - K(x_c)/g_0(x_c)} \tag{A.41}$$

in this expression the main difference with the *1D* case is the term with g_0, yet $r_0(t)$ is still a decreasing function of x_c as g_0 is an increasing function of x_c and $K(x_c)$ is a decreasing function ($C, D, A > 0$ and $J_0 < 0$), making the term $K(x_c)/g_0(x_c)$ a decreasing function of x_c.

The final expression we have obtained in Eq. A.41 for r_0 holds if several conditions can be verified, in particular if the total activity is one symmetric bump. We have not verified the symmetry condition, but numerically the shape of the bump seems symmetric. However, in the *2D* case it is common to observe small asymmetries in the shape of the bump.

A.2.2 Optimization of $v(\beta)$, *2D*

In this appendix we derive the expression of the asymmetry parameter β in terms of the input speed v, to maximize the total activity r_0. Considering the expression for the total activity for the population r_0 of units m obtained in Appendix A.2.1 for the *2D* case, this is equivalent to minimizing x_c (the size of the activity bump). To impose this we use the constraint of a single bump for the total input of activity,

$$AJ_1\sin(\beta-\psi+\phi)r_1 + D^2\sin(\phi-vt) = 0 \tag{A.42}$$

Using the expression we derived for r_1 in the *2D* case (analogous to r_0 in Eq. A.41) in Eq. A.42, we can derive and expression for the existence of the stable solution,

$$S' = \frac{J_0 g_0(x_c) + K(x_c)}{\sqrt{J_1^2 g_1^2 \cos(\Delta)^2 - 2J_1 g_1 \cos(\Delta)\cos(\Delta+\beta) + 1}} \tag{A.43}$$

As in *1D* this expression can be checked numerically for a given set of parameters as we do not know x_c. Yet, deriving by $\partial/\partial v$ and imposing $\partial x_c/\partial v|_{v=v_m} = 0$ we can find an expression for $v_m(\beta)$,

$$v_m = \frac{J_1 g_1(x_c) - 2\cos(\beta) \pm \sqrt{J_1^2 g_1^2(x_c) - 4J_1 g_1(x_c)\cos(\beta) + 4}}{2\tau \sin(\beta)} \tag{A.44}$$

where there is still a dependency with x_c expressed by $g_1(x_c)$. Checking the expression and choosing the appropriate solution $(-)$, we obtain the low $x_c \to \infty$ (or $J_1 g_1 \to \frac{1}{cos(\beta)}$) and high contrast $x_c \to 0$ (or $g_1 \to 0$) limits, re-obtaining the known solution from the literature in the *1D* for the first case, and the expression we obtained in Appendix A.1.3 for the high contrast expression,

$$v_m(\beta) = \frac{1 - \cos(\beta)}{\tau \sin(\beta)} \tag{A.45}$$

A.3 Numerical Simulations

The numerical simulation of Eq. 5.5 and Eq. 5.10 have been performed with the RK4 method (Runge-Kutta 4th order). Here we will give the parameters and discretization of the simulation in the *1D* case, and only the extra parameters for the *2D* case. First, we rewrite Eq. 5.1 as

$$\frac{\partial m(k,t)}{\partial t} = \frac{1}{\tau}(-m(k,t) + \left[\sum_{k'} w(k'-k)m(k',t)dk + C\left[1 - \epsilon + \epsilon\cos(k - vt)\right] - T\right]^+) \tag{A.46}$$

where we applied directly the RK4 method, see [PTVF92], to discretize the integral we use a trapezoidal rule. Here we use the parameters: $dt = 0.1$, $dk = 2\pi/60$, $\tau = .15$, $J_0 = -9.8$, $J_1 = 13.5$, $C = 5$, $T = 4.9$ and the total activity was computed over 1000 iterations. For the *2D* case the parameters were: $A = .2$, $dk = 2\pi/21$ and the total activity was computed over 100 iterations. Numerically, the system is quite robust to simulate and we were able to obtain similar results using the Euler method. Both implementations were performed in Matlab and are available from the author website.

B

Perspective projection

The camera we use to project 3D information onto a 2D plane uses a perspective projection, with intrinsic parameters: $f = 1.81$, $\tau = 2.41$ (associated to the wider aspect ratio we use, 800×600) and $n = -1$, $l = -10$ (near and far clipping distances), to obtain the 2D position using the 3D information following Eq. B.1,

$$\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & \tau & 0 & 0 \\ 0 & 0 & n & l \\ 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 & y_1 & z_1 & t_1 \\ x_2 & y_2 & z_2 & t_2 \\ x_3 & y_3 & z_3 & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{B.1}$$

where $(X, Y, Z)^T$ are the 3D coordinates in the world referential, approximately situated at the location of the subject, and $(x, y)^T$ in the camera plane referential. The matrix composed by x_i, y_i, z_i (rotation) and t_1, t_2, t_3 (translation) represents the rotation and translation from the world coordinate to the camera coordinates to align the z-axis of the camera and the world referential.

$$\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{bmatrix} 1.81 & 0 & 0 & 0 \\ 0 & 2.41 & 0 & 0 \\ 0 & 0 & -1 & -10 \\ 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1.0E+00 & -5.6E-04 & -3.1E-03 & 1.3E+00 \\ 1.2E-07 & 9.8E-01 & -1.8E-01 & -8.8E+01 \\ 3.1E-03 & 1.8E-01 & 9.8E-01 & -4.1E+02 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{B.2}$$

After the two referentials are aligned, the perspective projection represented by the parameters f, τ and n can be performed. In all our experiments, we use the values indicated in Eq. B.2.

C

PCA over the Optical flow

C.1 PCA Decomposition

The Principal Component analysis (PCA) is a process to transform a given space from a set of correlated variables into a (reduced) set of uncorrelated variables [CM98]. If we consider a set of m observations $\mathbf{x}_i$ each one of dimension n, the main idea is to retrieve a subset of c "components", such that any observation $\mathbf{x}_i$ can be described as,

$$\mathbf{x}_i = \sum_c \alpha_i \mathbf{v}_i \tag{C.1}$$

where α_i are scalar values and $\mathbf{v}_i$ represents each one of the components. When $c = n$, the PCA preserves dimensionality, but we use $c << n$, the case where we reduce dimensionality. In this case, there are several ways to obtain the $\mathbf{v}_i$ vectors. The chosen components are not guaranteed to be independent unless the data set is known to be jointly normal distributed. In the case where this hypothesis is not verified then the Independent Component Analysis (ICA) [Sap06] may be performed. The components selection criteria is not standard, but a common practice is to keep components in order to preserve more than 95% of the total variance, another criteria is to keep the components with an associated variance > 1, or until it is not significant to consider more components.

C.1.1 PCA using Single Value Decomposition (SVD)

In this work we use SVD to perform the PCA. PCA searches to identify the directions of more variation in the data, to express the same (or close) data with a few components or directions. The first step to perform the PCA is to subtract the mean of the observations, to center the data,

$$\mathbf{x}_i = \mathbf{x}_i - < \mathbf{x}_i > \tag{C.2}$$

where $< \mathbf{x}_i >$ is the mean value. The next step, is to compute the covariance matrix (the variance between each of the n components in the $\mathbf{x}_i$ vector), in matrix notation,

$$\mathbf{C} = \frac{1}{m}\mathbf{X}\mathbf{X}^T \tag{C.3}$$

where $\mathbf{X}$ is a matrix, in which each column corresponds to a vector $\mathbf{x}_i$, thus of size $n \times m$ (m is the number of observations). The covariance matrix ($\mathbf{C}$) can be obtained performing a SVD decomposition of the matrix $\mathbf{X}$, using the following idea. The SVD is a matrix decomposition valid for any matrix $\mathbf{B}$ of size $n \times m$, to write a matrix as the product of two orthonormal matrices ($\mathbf{U}$, $\mathbf{V}$) and one diagonal matrix $\mathbf{U}$, see Eq. C.4.

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{C.4}$$

To compute the SVD of $\mathbf{B}$, first the eigenvalues of $\mathbf{B}\mathbf{B}^T$ must be computed (an sorted in descending order), then its associated eigenvectors (columns in $\mathbf{V}$). Knowing these two matrices: $\mathbf{D}$ and $\mathbf{V}$, then $\mathbf{U}$ can be directly computed by using Eq. C.4. As the decomposition in Eq. C.4 can be applied to any matrix, in particular, it can applied to $\mathbf{X}$, to write $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, writing the covariance matrix as,

$$\mathbf{C} = \frac{1}{m}\mathbf{U}\mathbf{D}^2\mathbf{U}^T \tag{C.5}$$

The values in $\mathbf{D}^2$ correspond to the variance associated to a particular component of the matrix $\mathbf{V}$, and as the SVD sort the values, they are arranged in descending order. This order is particularly important as the first value in $\mathbf{D}^2$ represents the highest value associated to the first principal component, which can be found in the first column of $\mathbf{U}$.

C.1.2 PCA and image rotation

A relevant aspect of the PCA when computed over images is the effect of a rotation (of the image) over the principal components. Lets suppose that each vector $\mathbf{x}$ represents an image by rows, so a 800×600 image gives a vector of size $n = 480000$, and that we perform a rotation around the center of the image,

$$\mathbf{x}_i' = \mathbf{R}\mathbf{x}_i \tag{C.6}$$

here the matrix $\mathbf{R}$ corresponds to the rotation. In $\mathbf{R}$ there is a single 1 per row (as the pixels will be swaped in a rotation), where we ignore effects due to the rectangular shape of the image and to discretization. When many samples are used, and if each image goes through the same rotation, we can write,

$$\mathbf{X}' = \mathbf{R}\mathbf{X} \tag{C.7}$$

Following the computation of the covariance matrix, we can write the covariance matrix for the rotated set of images as in Eq. C.8.

$$\mathbf{C}' = \frac{1}{m}\mathbf{R}\mathbf{X}\mathbf{X}^T\mathbf{R}^T \tag{C.8}$$

Eq. C.8 can we written as $\mathbf{C}' = \mathbf{R}\mathbf{C}\mathbf{R}^T$ using Eq. C.5, and thus if we write it in terms of the SVD,

$$\mathbf{C}' = \frac{1}{m}(\mathbf{R}\mathbf{U})\mathbf{D}^2(\mathbf{R}\mathbf{U})^T \tag{C.9}$$

We can interpret Eq. C.9 as a rotation of the principal components, by consequence, a rotation of each image produces a rotation over the principal components. The result is not surprising, as geometrically, the PCA can be seen as a rotation to maximize the covariance, thus a rotation of the original data, will also rotate the principal components. As a result, the PCA is not invariant to rotations.

C.1.3 PCA and image scaling

The same demonstration can be performed if we study a "zoom-out" over an image. This operation corresponds to keeping some pixels in the image, and setting the other ones to zero, and we represent it in a matrix **S**. We notice that it is a simplification, because the "zoom-out" operation requires also to perform a low-pass filter to avoid artifacts, thus changing the original image. The "zoom-out" over the image can be written in matrix notation as,

$$\mathbf{X}' = \mathbf{S}\mathbf{X} \tag{C.10}$$

where **S** has a single 1 per row, and some rows are filled only with zeros (some pixels are ignored). Making the analogy with the rotation, the associated covariance matrix over the sub-sampled set of images can be written as,

$$\mathbf{C}' = \frac{1}{m}(\mathbf{S}\mathbf{U})\mathbf{D}^2(\mathbf{S}\mathbf{U})^T \tag{C.11}$$

We have shown that the change of scale, over a set of images, can approximately be seen as a sub-sampling of the principal components. Therefore, the PCA is not invariant to sub-sampling (zoom-in/out) over images, as the components will be subsampled in a similar way.

C.2 Application to the optical flow

We apply now the PCA to the optical flow, to look for the most relevant (representative of most of the variance) local patterns of motion. At the same time, we verify if the analysis is independent of the kind of sequence and the optical flow extraction technique.

In order to compute the PCA, we identify the observation vectors as the optical flow in a small patch. Considering a square patch size of side L (in pixels), the dimension n of the observation is $n = 2L^2$, where the factor 2 comes from the velocity vector (speed and direction). We build patches with 50% overlapping receptive fields. In our analysis, the overlapping will not change the dimension but the number of available examples, only the receptive field size (L) will change this dimension.

The experiments we present were applied to three databases: the ICCV database [MNCG01], the KTH database [SLC04] and our own human motion database (Loria). The first database contains a sequence of rigid moving objects where the "perfect" optical flow is available. The KTH database consists in a series of human actions (to box, to clap, to wave), with different actors and cameras where the ground-truth optical flow is not available. Our database consists in the animation of three human motions, the same as the KTH database, but as animations (less noisy than KTH). The main objective of the experiments is to determine if local optical flow patterns are significant descriptors of these sequences, their shapes and if different detectors are necessary for each kind of sequence. Also, we search to determine if this decomposition depends on the optical flow extraction technique.

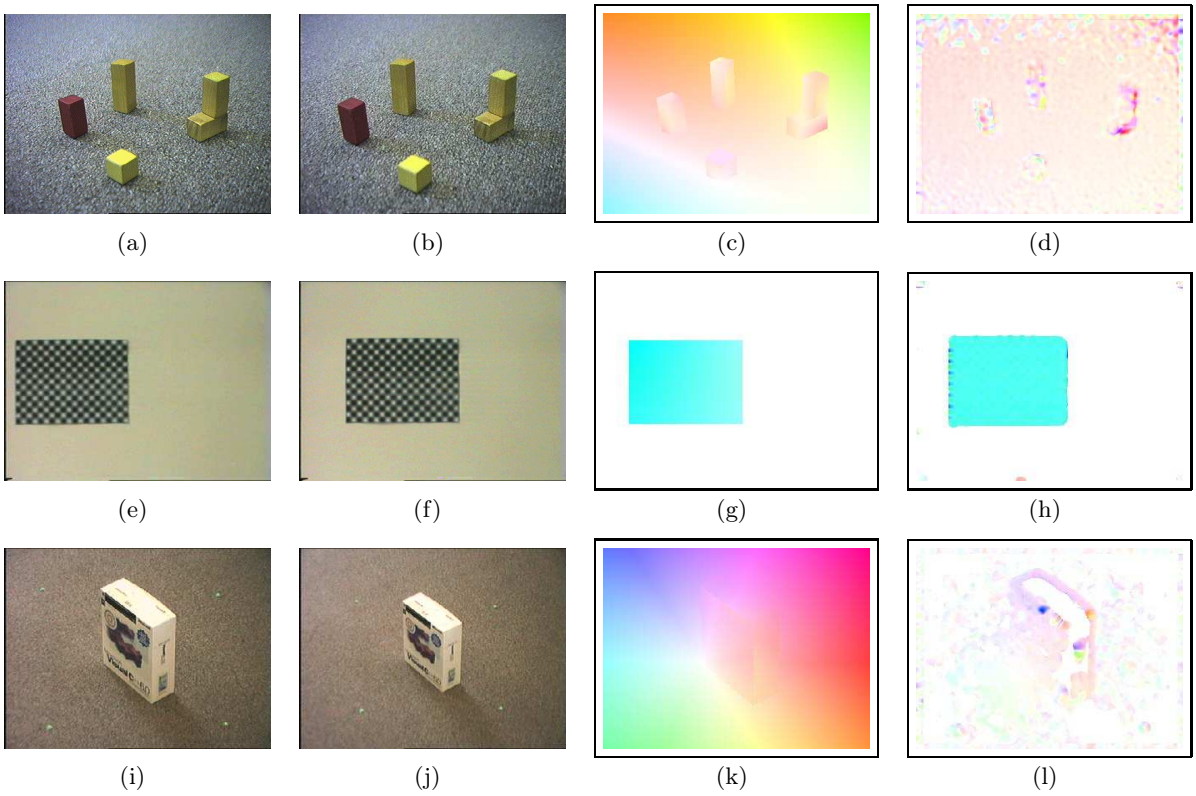


FIGURE C.1 – In the first and second column, two frames of the block sequence in the ICCV DB, the blocks remains static and the camera moves. The optical flow is in the Middlebury color code [BRS+07] (color represents orientation and saturation represents speed), using ground-truth (third column) and the multi-scale Lukas & Kanade algorithm (fourth column).

C.2.1 Rigid movements (ICCV database)

The ICCV database provides ground-truth optical flow, that we call OFG, using three objects: a check-board, a box and wood blocks (see Figure C.1), which dimensions are known, so that the optical flow can be predicted, see [MNCG01] for more details.

We start by comparing the PCA extraction using the available "perfect" optical flow and the one we extract by using the Lucas & Kanade [Luc85, BK08] algorithm in its serial multi-scale version (see Chapter 3), that we call OFR. The size of the patches was 14 pixels, here we follow the anatomical references indicated by [CG05]. The main observation in this experiment is that Lucas & Kanade algorithm gives more noisy results. Despite this, the firsts PCA components are similar in the sense that in both cases components represents mainly discontinuities in the flow. We also note, that the first two components are very similars, see Figure C.3.

C.2.2 Human Motion (KTH database)

In the KTH database, a group of actions are performed by different actors (see Figure C.2). This database is quite noisy, as it is available as compressed video, and it was captured at 25Hz. We repeated the same analysis over the optical flow for the ICCV database, using the same parameters. But this time, only over the computed optical flow using the Lucas & Kanade

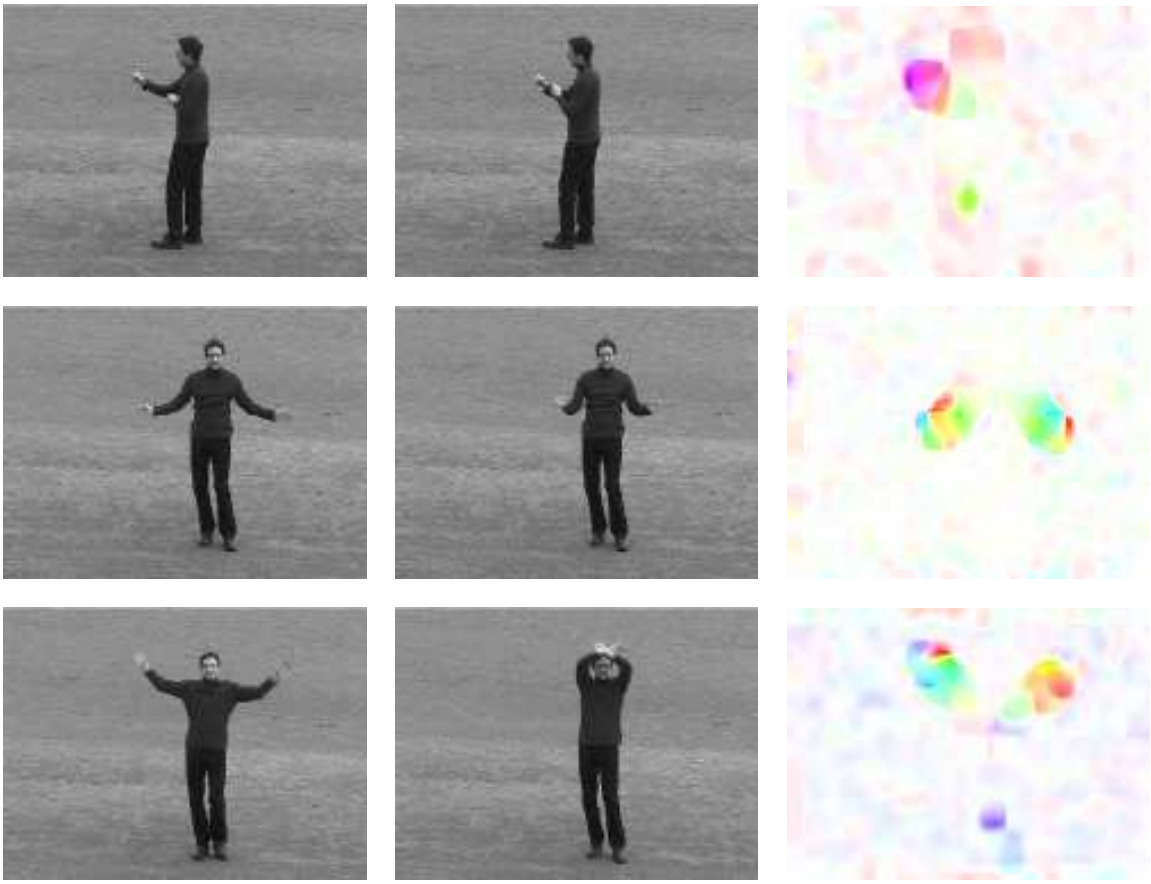


FIGURE C.2 – The 3 sequences of the KTH DB, (camera remains static). In each row: two non-consecutives frames and the optical flow in the Middlebury color code [BRS+07] (color represents orientation and saturation represents speed), using the multi-scale Lukas & Kanade algorithm.

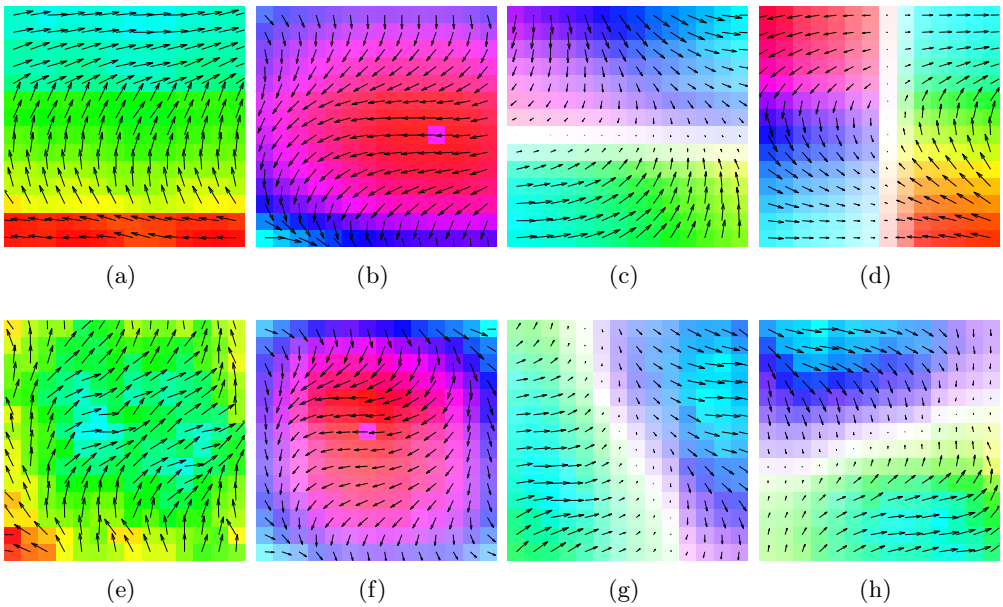


FIGURE C.3 – First 4 PCA components for the ground-truth (first row) and the computed optical flow (second row). In both cases, the size of the operator was 14 pixels with a 50% overlap, using the optical flow from 18 frames, in the 3 sequences: blocks, grid and box, see Figure C.1.

algorithm, since the ground-truth information is not available in this database. Once the optical flow was computed, then the PCA was performed over the three sequences simultaneously over 18 frames. The first 4 components can be seen in the first row of Figure C.4.

C.2.3 Human motion (Loria database)

In the Loria database, that we built, a test subject performs a set of 3 actions. The actions were "to clap", "to fight" and "to wave", see the second row of Figure C.4. Once the sequences were captured (in 3D), we animate these data using a mesh model of the body. Compared to the KTH database, the movements are similar, but the animation has much less noise, as there is no associated compression in the capture. We compute the optical flow in this database from the images, so the perturbations commonly associated to the optical flow extraction can be observed, yet the signal is not as noisy as in the KTH database. The frame rate in this database is 100Hz compared to the 25Hz in the KTH database. Once the optical flow was computed, then the PCA was performed over the three sequences simultaneously over 18 frames. The first 4 components can be seen in the second row of Figure C.4.

C.2.4 Results

The results of the experiments show that local patterns computed over the optical flow are invariant to the optical flow extraction technique, as we explain in C.2.1, and illustrate in Figure C.3. So, even if optical flow extraction techniques are noisy, this does not affect the PCA decomposition. This result can be retrieved, not only for simple sequences as in the ICCV database but also over sequences of human motion either real or animated, where the PCA components are not very different, see Figure C.4.

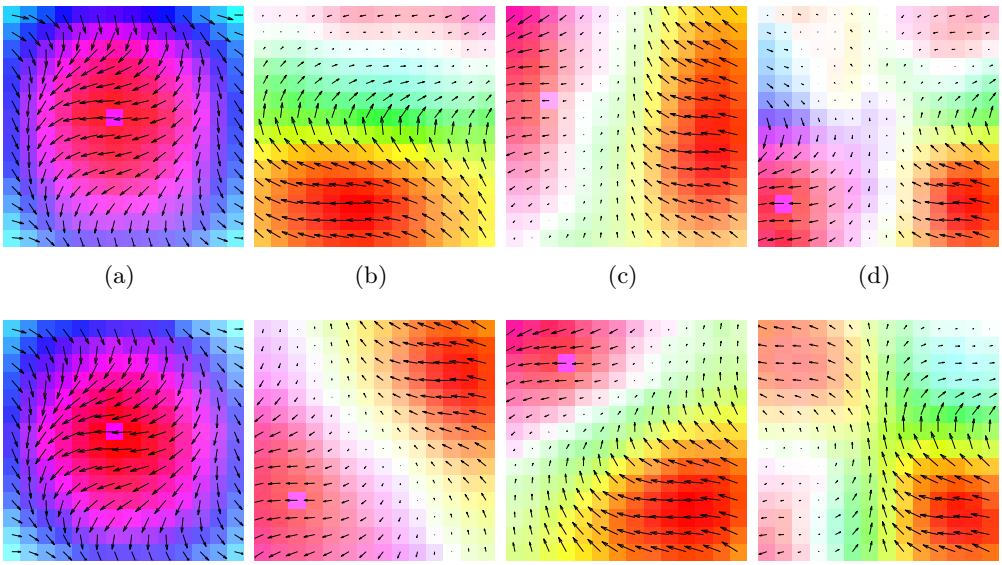


FIGURE C.4 – The first four PCA components for the KTH (first row) and the Loria (second row) database. In both databases, there are 3 sequences, the size of the operator was 14×14 pixels with a 50% overlap, using the optical flow from 18 frames. The actions in both database are the same, with the exception of "to fight" where the viewpoint is different.

TABLE C.1 – Percentage of the total variance at 1, 7 and 49 components, for the four databases.

PCA Comp.	Database	% of Total Variance
1	KTH	31%
	OFG	84%
	OFR	57%
	Loria	21%
7	KTH	65%
	OFG	99%
	OFR	81%
	Loria	57%
49	KTH	89%
	OFG	99%
	OFR	94%
	Loria	85%

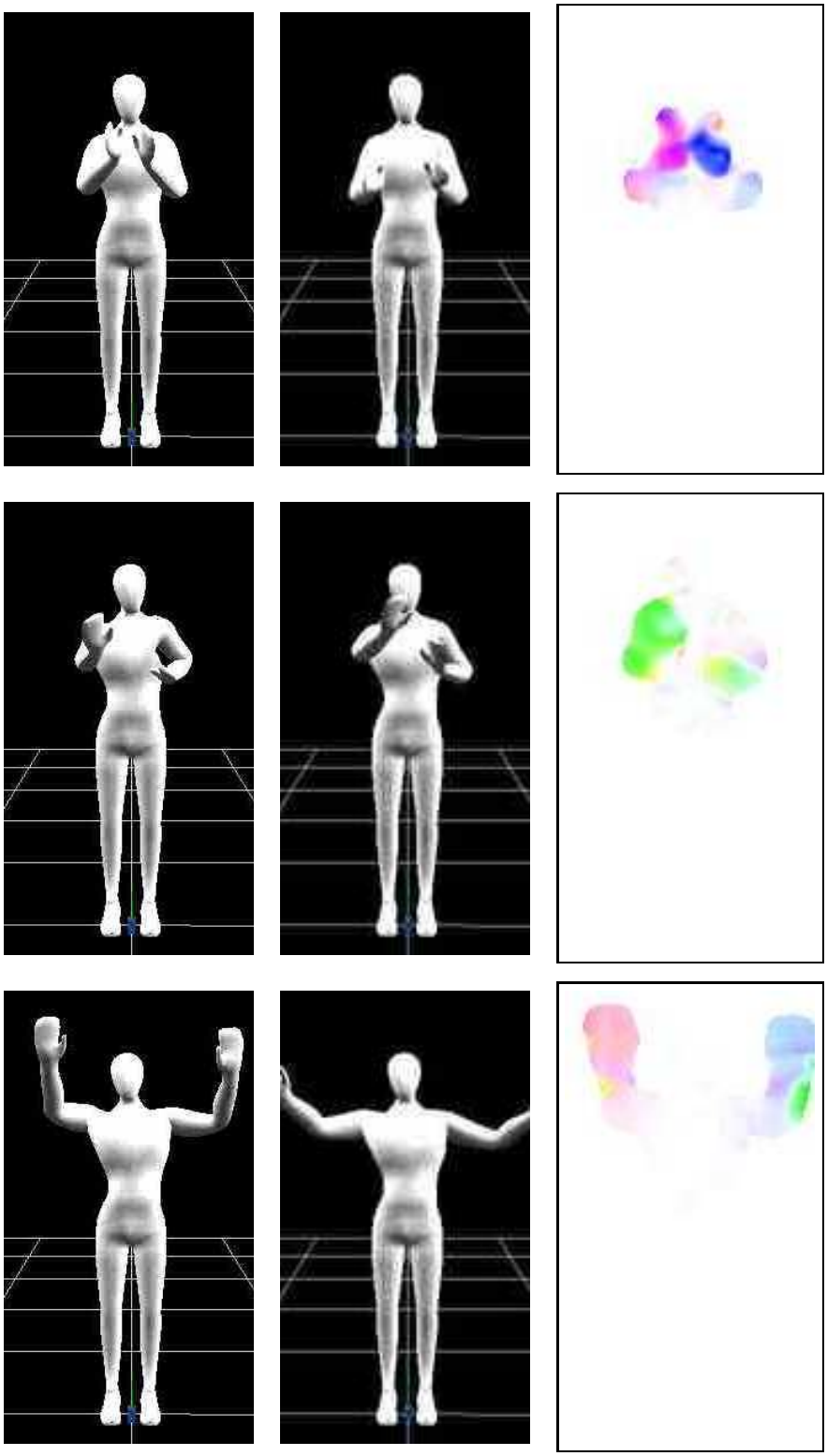


FIGURE C.5 – The 3 sequences of our database (animated from the 3D capture). In each row: two non-consecutives frames and the optical flow in the Middlebury color code [BRS+07] (color represents orientation and saturation represents speed), using the multi-scale Lukas & Kanade algorithm.

As for the number of components necessary to describe the sequences, this depends greatly on the sequence. For the simple ICCV database, already 1 component represents 84% of the variance, see Table C.1. In more complicated sequences (KTH, Loria), the first component represents only 21%-31%, and around 7 components are necessary to account for 60% of the total variance in these sequences, see Table C.1. We notice, however, that some components in these sequences seem to differ by a rotation, see components (b) and (c) in Figure C.4. This may explain why in human motion sequences we require more components to describe the total variance, as the movements is in many directions and as the PCA is not invariant to rotation, similar components are extracted at different orientations. Finally, we notice a very interesting component, see Figure C.3(b), Figure C.3(f) and Figure C.4(a), Figure C.4(e), that appears in all our experiments either as the most of the second most relevant local flow pattern.

D

Facial motion discrimination

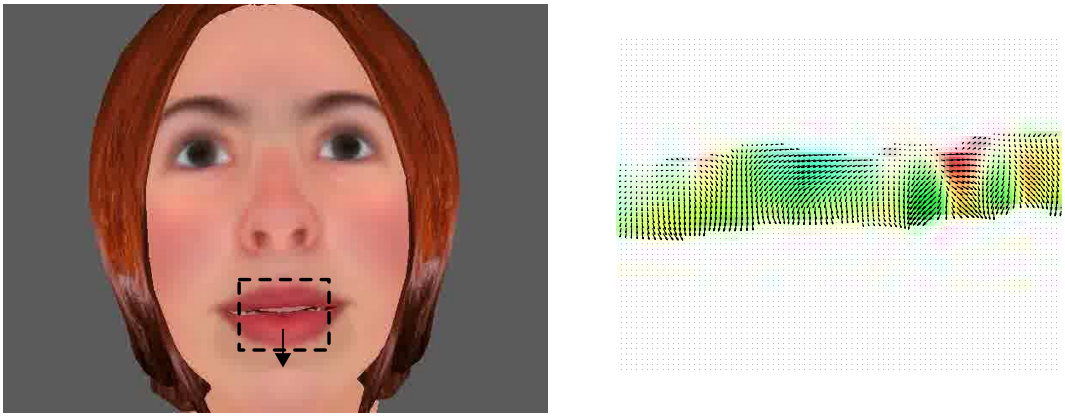
D.1 Face movements

We have performed an exploratory work about the classification of facial movements. Our objective was to verify if the proposed ACNFT model is able to distinguish this kind of movements. The main difficulties with face movements are the short times involved and the lack of spatially localized movements near the mouth, where almost every pixel moves (see Figure D.1). The lack of spatial structure made difficult to extract few relevant spatial trajectories, as the ACNFT requires to encode a spatiotemporal sequence.

Instead of classifying face videos directly, we evaluate our ACNFT model using face-related spatially localized parameters. More precisely, we chose to study mouth opening and mouth width along time. Even though mouth opening/width are simple parameters, our acquisition setup was not adapted for this experiment (it requires close-distance lenses and more than 4 high speed cameras). We decided to study a very different approach, generating face animations from audio signal [LT11], that can be more easily captured. To explore this approach, we have participated in the development of an audio-driven face animation approach [CN10], as part of a joint effort in the context of a STIC Amsud Program [14]. To validate this approach in terms of the synthesized animation, we performed perceptive evaluation to validate the speech-driven animation [LT11], achieving good perceptive quality.

Thanks to the speed-driven animation, we were able to generate a simple digit database composed by 10 different face gestures. The two parameters trajectories, see Figure D.2, taken together, characterize the input to perform the discrimination over time. In the illustration, it can be noticed that the trajectories are generally closed curves (or near that), because the mouth movements start and end at the rest position. For example, the digit 8 (phonetically in Spanish /o̤ ∫ o̤/) involves mouth opening and closing almost symmetrically, where most of the sound is produced inside the vocal cavity.

14. Collaboration Program between the INRIA and scientific partners in South America. We participate in the project BAVI with partners in Santiago (Chili) and Rosario (Argentine).



(a) (b)

FIGURE D.1 – (a) Animated character, pronouncing the beginning of the digit zero, the inferior lip is moving down. (b) Zoom in the rectangle area in (a), all the lip is moving down.

D.2 Face movements: classification using ACNFT

The database consists of one subject with normal speech abilities, pronouncing the digits from 0 to 9 in Spanish, 3 times each. In order to have the same duration, each sound was captured with the same time window and aligned at the beginning of the interval. Using the audio-visual associative model [LT11], the visual parameters of mouth width and opening were obtained from audio information directly.

D.2.1 Experiments

We study two scenarios, the first where we study the classification performance of the ACNFT model by learning from the average of all trials and testing with one of the samples. The second setup consisted in learning with one trial, and evaluating the classification by using a different trial. As the trajectory information is normalized, we map this information into a larger space (200×200) to allow the ACNFT to encode the information. The simulation was performed using the RK4 method for the differential equation (more details in appendix A.3).

Mean trajectory learning

The experimental results indicate that, in general, the digits can be recognized $(2, 3, 5, 6, 9, 10)$, looking by rows in Figure D.3(a). There are some digits that are quite similar in the space in which we represent facial movements, like 2 and 4 (see Figure D.2). The similarity can be retrieved again in the classification with the ACNFT, for example, in the second row of Figure D.3(a), where 2 and 4 generate high responses.

Single trial learning

To train the ACNFT with a single trial also gives interesting results. Compared with the first experience, the populations in Figure D.3(b) are more active in general. The presence of more (than in the mean trajectory training case) noise could explain this effect, because the modified kernels are more sparsely located in the map, thus more units can be potentially activated. In

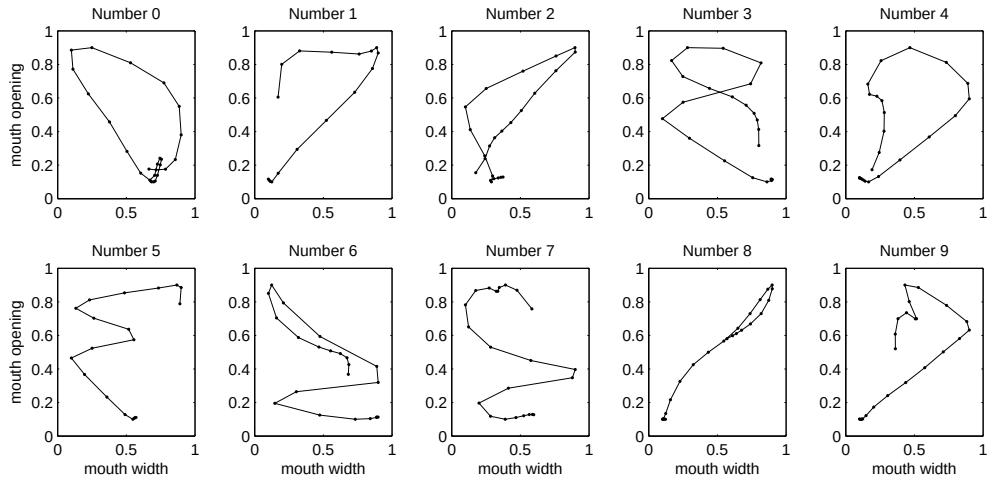
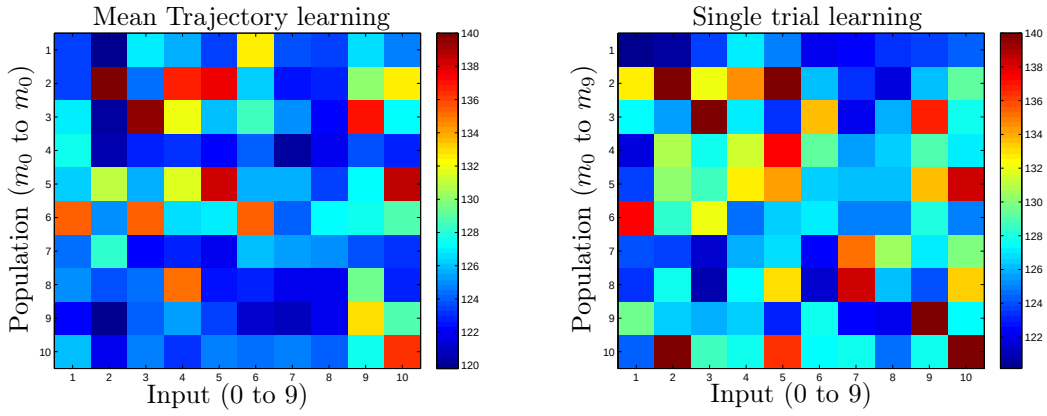


FIGURE D.2 – Ten different digits, using mouth opening and width as parameters. The data correspond to the average movement of one subject, over 3 trials.



(a) (b)

FIGURE D.3 – (a) Confusion matrix, where the mean trajectory is used for the learning stage. (b) Confusion matrix, when a single trial is used for the learning stage.

terms of discrimination, single trial learning delivers lower performance than using the average trajectory, i.e. along a row in Figure D.3(b) there are more locations where the activity is high. However, the observed error is related to the confusion of classes (the right class still has a high value), for example, for row number 2, there is a high activity at 2 (the right answer) but also at 5 (wrong classification).

D.2.2 Discussion

We have presented the application of the ACNFT model, for the classification of facial gestures, specifically the pronounciation of 10 digits from 0 to 9 in Spanish. The results of our evaluation highlights once again that the features to be encoded by the ACNFT need to be decomposable into a set of spatiotemporal trajectories. To this purpose, we use artificial parameters: mouth opening and mouth width, and we evaluate our model with these features. The result are promising, from single trial learning some discrimination can already be delivered, and when the learning takes into account more example the classification improves (mean trajectories). Facial gestures are a difficult case of study, because some sounds simply cannot be disambiguated only using mouth information. Tongue-related sounds, like /o̤ ∫ o̤/), require more information. Even in this context our ACNFT model manages to deliver discrimination between the sequences, though it delivers simultaneous high activities for several populations.

E
Publications

ECVP 2011. Mauricio Cerda and Bernard Girau. "Motion decomposition for biological and non-biological movements" (abstract). European conference on visual perception, Toulouse, France.

INRIA Research Report. Mauricio Cerda and Bernard Girau. "Spatiotemporal pattern coding using Neural Fields: Optimal parameter estimation". Research Report RR-7543, INRIA, Février 2011.

ICME 2011. Lucas Terissi, Mauricio Cerda, Juan C. Gomez, Nancy Hitschfeld-Kahler, Bernard Girau, Renato Valenzuela. "Animation of generic 3D head models driven by speech". IEEE International Conference on Multimedia and Expo, Juillet 11 - 15, Barcelona, Spain.

SCCC 2010. Mauricio Cerda, Renato Valenzuela and Nancy Hitschfeld-Kahler, « Generic face animation ». XXIX International Conference of the Chilean Computer Society, Novembre 15 - 19, Antofagasta, Chile.

BICS 2010. Mauricio Cerda and Bernard Girau. "A bio-inspired neural model to discriminate-visual sequences". Brain Inspired Cognitive Systems, Juillet 14 - 16, Madrid, Spain.

BIONETICS 2009. Mauricio Cerda, Lucas Terissi and Bernard Girau. "Bio-inspired speed detection and discrimination. 4th International Conference on Bio-Inspired Models of Network, Information, and Computer Systems", Décembre 9 - 12, Avignon, France.

NeuroComp 2009. Mauricio Cerda and Bernard Girau. « Visual pattern classification by neural fields ». Troisième conférence française de Neurosciences Computationnelles, Septembre 16 - 18, Bordeaux, France.

ReConfig 2008. Hugo Barron-Zambrano, Cesar Torres-Huitzil and Mauricio Cerda. "Flexible Architecture for Three Classes of Optical Flow Extraction Algorithms". In International Conference on ReConFigurable Computing and FPGAs, Décembre 3 - 5. Cancun, Mexico.

ESANN 2008. Mauricio Cerda and Bernard Girau. "A neural model with feedback for robust disambiguation of motion". In European Symp. on Artificial Neural Networks, Avril 23 - 25 Bruges, Belgium.

Glossaire

Glossaire avec les principaux acronymes et concepts utilisés dans le manuscrit.

ACNFT : Asymmetric continuum neural field theory.
ACP : Analyse en Composantes Principales.
AEI : Asymmetric Excitatory-Inhibitory (model).
ANOVA : ANalysis Of VAriance.

CNFT : Continuum neural field theory.
CVPR : Computer Vision & Pattern Recognition Conference.

EBA : Extrastriate Body Area.
ERD : Elaborated Reichardt Detector.

F5 : F5 area (ventral premotor cortex).
FBA : Fusiform Body Area.
FFA : Fusiform face area.
fMRI : Functional Magnetic Resonance Image.
FPGA : Field Programable Gateway Array.

GPU : Graphical Processing Unit.

HSV : Hue Saturation Value.

ICA : Independent Components Analysis.
ICCV : International Conference Computer Vision.
IOC : Intersection Of Constraints.
IRM : Image Resonance Magnetic.
IT : Inferotemporal area.

KO : Kinetic Occipital area.
KTH : Royal Institute of Technology (Sweden).

LGN : Lateral Geniculate Nucleus.
LISP : List Processing (functional programming language).

MST : Medial Superior Temporal area.
MT : Middle Temporal Area.

OFA : Occipital Face Area.
OFG : Optical Flow Ground-truth.
OFR : Optical Flow Real (from Lucas & Kanade algorithm).
OLS : Ordinary Least Squares.

OpenCV : Open Computer Vision (software library).

PCA : Principal Components Analysis.
PL : Point-light stimuli.

RGB : Red Green Blue.

SAD : Sum of Absolute Difference.
SE : Symmetric Excitatory (model).
STS : Superior Temporal Sulcus.
SVD : Single Value Decomposition.

V1 : Visual area 1.
V2 : Visual area 2.
V3 : Visual area 3.
V4 : Visual area 4.

References

[AB85] Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. AM. A*, 2(2):284–299, 1985.

[AL00] Cowey A. and Vaina L.M. Blindness to form from motion despite intact static form perception and motion detection. *Neuropsychologia*, 38(13):566–578, 2000.

[Ama77] Shun-Ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

[AS08] Dimitrios S. Alexiadis and George D. Sergiadis. Narrow directional steerable filters in motion estimation. *Comput. Vis. Image Underst.*, 110(2):192–211, 2008.

[BB05] Richard T. Born and David C. Bradley. Structure and function of visual area mt. *Annual Review of Neuroscience*, 28(1):157–189, 2005.

[BBS98] I. Bülthoff, H.H. Bülthoff, and P. Sinha. Top-down influences on stereoscopic depth-perception. *Nature Neuroscience*, 1, No. 3:254–257, 1998.

[BD02] Gary R. Bradski and James W. Davis. Motion segmentation and pose recognition with motion history gradients. *Mach. Vision Appl.*, 13:174–184, July 2002.

[BFBB94] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. *CVPR*, 92:236–242, 1994.

[BFM10] Fréderic V. Barthélemy, Jérome Fleuriet, and Guillaume S. Masson. Temporal dynamics of 2d motion integration for ocular following in macaque monkeys. *Journal of Neurophysiology*, 103:1275–1282, March 2010.

[Bis84] P. O. Bishop. *Handbook of Physiology, The Nervous System, Sensory Processes*, chapter Processing of Visual Information within the Retinostriate System, pages 341–424. John Wiley & Sons, Inc., 1984.

[BK08] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 1st edition, October 2008.

[BL98] Marc Bolduc and Martin D. Levine. A review of biologically motivated space-variant data reduction models for robotic vision. *Comput. Vis. Image Underst.*, 69:170–184, February 1998.

[Bla92] M. J. Black. *Robust incremental optical flow*. PhD thesis, Yale University, New Haven, USA, 1992.

[BN04] P. Bayerl and H. Neumann. Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16(10):2041–2066, 2004.

[BN07a] Pierre Bayerl and Heiko Neumann. Disambiguating visual motion by form-motion interaction - a computational model. *International Journal of Computer Vision*, 72:27–45, 2007. 10.1007/s11263-006-8891-8.

[BN07b] Pierre Bayerl and Heiko Neumann. Disambiguating visual motion by form-motion interaction–a computational model. *Int. J. Comput. Vision*, 72:27–45, April 2007.

[BP94] Bennett I Bertenthal and Jeannine Pinto. Global Processing of Biological Motions. *Psychological Science*, 5(4):221–225, 1994.

[BRS+07] Simon Baker, Stefan Roth, Daniel Scharstein, Michael J. Black, J.P. Lewis, and Richard Szeliski. A database and evaluation methodology for optical flow. *Computer Vision, IEEE International Conference on*, 0:1–8, 2007.

[BS07] R. Blake and M. Shiffrar. Perception of human motion. *Annu. Rev. Psychol.*, 58:47–73, 2007.

[CBK03] K.M.G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–77 – I–84 vol.1, june 2003.

[CCPB03] Kevin R Duffy Christopher C Pack, Margaret S Livingstone and Richard T Borna. End-stopping and the aperture problem: Two-dimensional motion signals in macaque v1. *Neuron*, Volume 39(4):671–680, August 2003.

[CG05] Antonino Casile and Martin A. Giese. Critical features for the recognition of biological motion. *J. Vis.*, 5(4):348–360, 4 2005.

[CG08] Mauricio Cerda and Bernard Girau. A neural model with feedback for robust disambiguation of motion. In *ESANN*, pages 505–510, 2008.

[CGM98] J. Chey, S. Grossberg, and E. Mingolla. Neural dynamics of motion processing and speed discrimination. *Vision Research*, 38(18):2769–2786, September 1998.

[CH10] Beck C. and Neumann H. Interactions of motion and form in visual cortex–a neural model. *Journal of Physiology Paris*, 104:61–70, 2010.

[CHKM07] Mauricio Cerda, Nancy Hitschfeld-Kahler, and Domingo Mery. Robust tree-ring detection. In Domingo Mery and Luis Rueda, editors, *PSIVT*, volume 4872 of *Lecture Notes in Computer Science*, pages 575–585. Springer, 2007.

[CM98] Vladimir S. Cherkassky and Filip Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., New York, NY, USA, 1998.

[CN10] Valenzuela R. Cerda, M and Hitschfeld-Kahler N. Generic face animation. In *XXIX International Conference of the Chilean Computer Society (SCCC)*, pages 252–257. IEEE Computer Society, 2010.

[Coo05] S. Coombes. Waves, bumps, and patterns in neural field theories. *Biological Cybernetics*, 93:91–108, 2005. 10.1007/s00422-005-0574-y.

[Cre93] Daniel Crevier. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc., New York, NY, USA, 1993.

[CVP05] *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. IEEE Computer Society, 2005.

[CW03] Leo M. Chalupa and John S. Werner. *The Visual Neurosciences*. MIT Press, Cambridge, MA, 2003.

[Der87] Rachid Deriche. Using canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1:167–187, 1987. 10.1007/BF00123164.

[DOTG91] Iii Davis, Sylvia Ounpuu, Dennis Tyburski, and James R. Gage. A gait analysis data collection and reduction technique. *Human Movement Science*, 10(5):575–587, October 1991.

[DR06] Philip DeCamp Rony Kubat Michael Fleischman Brandon Roy Nikolaos Mavridis Stefanie Tellex Alexia Salata Jethran Guiness Michael Levit Peter Gorniak Deb Roy, Rupal Patel. The human speechome projec. In *Proceedings of the Twenty-eighth Annual Meeting of the Cognitive Science Society (CogSci)*, 2006.

[DT02] J.W. Davis and S.R. Taylor. Analysis and recognition of walking movements. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 315 – 318 vol.1, 2002.

[Duf98] Charles J. Duffy. Mst neurons respond to optic flow and translational movement. *Journal of Neurophysiology*, 80(4):1816–1827, 1998.

[DW95] CJ Duffy and RH Wurtz. Response of monkey mst neurons to optic flow stimuli with shifted centers of motion. *The Journal of Neuroscience*, 15(7):5192–5208, 1995.

[EMVK09] Maria-Jose Escobar, Guillaume Masson, Thierry Vieville, and Pierre Kornprobst. Action recognition using a bio-inspired feedforward spiking network. *International Journal of Computer Vision*, 82:284–301, 2009. 10.1007/s11263-008-0201-1.

[FB83] B. Fischer and R. Boch. Saccadic eye movements after extremely short reaction times in the monkey. *Brain Research*, 260(1):21 – 26, 1983.

[FB05] Stefanos E. Folias and Paul C. Bressloff. Stimulus-locked traveling waves and breathers in an excitatory neural network. *SIAM Journal of Applied Mathematic*, 65(6):2067–2092, 2005.

[FC07] G.D. Field and E.J. Chichilnisky. Information processing in the primate retina: Circuitry and coding. *Annual Review of Neuroscience*, 30(1):1–30, 2007.

[FL97] Vincent P. Ferrera and Stephen G. Lisberger. Neuronal responses in visual areas mt and mst during smooth pursuit target selection. *Journal of Neurophysiology*, 78(3):1433–1446, 1997.

[FM00] David Ferster and Kenneth D. Miller. Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience*, 23:441–471, 2000.

[FTL09] Winrich A. Freiwald, Doris Y. Tsao, and Margaret S. Livingstone. A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12(9):1187–1196, August 2009.

[Gib50] J. J. Gibson. *The perception of the visual world.* Greenwood Pub Group, 1950.

[GKL97] Karl R. Gegenfurtner, Daniel C. Kiper, and Jonathan B. Levitt. Functional properties of neurons in macaque area v3. *Journal of Neurophysiology*, 77(4):1906–1923, 1997.

[GM09] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2009.

[GP01] Herbert Goldstein and Charles P. Poole. *Classical Mechanics.* Addison Wesley, June 2001.

[GP03] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, March 2003.

[Hee87] David J. Heeger. Model for the extraction of image flow. *J. Opt. Soc. AM. A*, 4(8):1455–1471, 1987.

[HhYW05] Gang Hua, Ming hsuan Yang, and Ying Wu. Learning to estimate human pose with data driven belief propagation. In *CVPR*, pages 747–754, 2005.

[HJG+01] Jean-Michel Hupe, Andrew C. James, Pascal Girard, Stephen G. Lomber, Bertram R. Payne, and Jean Bullier. Feedback connections act on the early part of the responses in monkey visual cortex. *J Neurophysiol*, 85(1):134–145, 2001.

[HK] Ralph Nelson Helga Kolb, Eduardo Fernandez. Webvision, the organization of the retina and visual system. `http://webvision.med.utah.edu/`.

[HK87] E C Hildreth and C Koch. The analysis of visual motion: From computational theory to neuronal mechanisms. *Annual Review of Neuroscience*, 10(1):477–533, 1987.

[HL87] DH Hubel and MS Livingstone. Segregation of form, color, and stereopsis in primate area 18. *The Journal of Neuroscience*, 7(11):3378–3415, 1987.

[HP00] Harold Hill and Frank E. Pollick. Exaggerating Temporal Differences Enhances Recognition of Individuals from Point Light Displays. *Psychological Science*, 11(3):223–228, 2000.

[HPL+98] J. M. Hupe, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394:784–787, 1998.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *ARTIFICAL INTELLIGENCE*, 17:185–203, 1981.

[HS98] David Hansel and Haim Sompolinsky. *Methods in Neuronal Modeling: From synapses to networks*, chapter Modeling Feature Selectivity in Local Cortical Circuits. MIT Press, Cambridge, MA, USA, 1998.

[HVE00] Jay Hegdé and David C. Van Essen. Selectivity for complex shapes in primate visual area v2. *The Journal of Neuroscience*, 20(5):RC61, 2000.

[HW62] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[HW68] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[IF01] S. Ioffe and D.A. Forsyth. Human tracking with mixtures of trees. pages I: 690–695, 2001.

[JJT91] Michael R. M. Jenkin, Allen D. Jepson, and John K. Tsotsos. Techniques for disparity measurement. *CVGIP: Image Understanding*, 53(1):14 – 30, 1991.

[Joh73] G. Johansson. Visual perception of biological motion and model for its analysis. *Percept. Psychophys.*, 14:195–204, 1973.

[JR02] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *Int. J. Comput. Vision*, 46:81–96, January 2002.

[JS04] Alissa Jacobs and Maggie Shiffrar. Walking perception by walking observers. *J. Vis.*, 4(8):218–218, 8 2004.

[JSWP07] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8, 2007.

[Jun06] Gregory Junker. *Pro OGRE 3D Programming (Pro).* Apress, Berkely, CA, USA, 2006.

[KHM00] I. A. Karaulova, P. M. Hall, and A. D. Marshall. A hierarchical model of dynamics for tracking people with a single video camera. In *in: British Machine Vision Conference*, pages 352–361, 2000.

[Koe86] Jan J. Koenderink. Optic flow. *Vision Research*, 26(1):161 – 179, 1986.

[KRW90] M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten. Measurement of lower extremity kinematics during level walking. *Journal of Orthopaedic Research*, 8(3):383–392, 1990.

[KSJ00] Eric Kandel, James Schwartz, and Thomas Jessell. *Principles of Neural Science.* McGraw-Hill Medical, 4th edition, July 2000.

[Kvv85] J. J. Koenderink, A. J. van Doorn, and W. A. van de Grind. Spatial and temporal parameters of motion detection in the peripheral visual field. *Journal of the Optical Society of America A*, 2:252–259, February 1985.

[KY06] Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B*, 361:2109–2128, 2006.

[Lap05] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64:107–123, 2005.

[Lee11] Sangyoon Lee. Motionviewer, a small visualization program to preview motion data, 2011. `http://www.evl.uic.edu/sjames/mocap/motionviewer.html`.

[LKM94] J. B. Levitt, D. C. Kiper, and J. A. Movshon. Receptive fields and functional architecture of macaque v2. *Journal of Neurophysiology*, 71(6):2517–2542, 1994.

[LN03] J. Liu and W. T. Newsome. Functional Organization of Speed Tuned Neurons in Visual Area MT. *J. Neurophysiol.*, 89(1):246–256, 2003.

[LPB01] Margaret S. Livingstone, Christopher C. Pack, and Richard T. Born. Two-dimensional substructure of mt receptive fields. *Neuron*, 30(3):781 – 793, 2001.

[LT11] Juan C. Gomez Nancy Hitschfeld-Kahler Bernard Girau Renato Valenzuela Lucas Terissi, Mauricio Cerda. Animation of generic 3d head models driven by speech. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2011.

[Luc85] Bruce David Lucas. *Generalized image matching by the method of differences.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 1985.

[MAGN85] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome. The analysis of moving visual patterns. In C. Chagas, R. Gattass, and C.Gross, editors, *In Pattern Recognition Mechanisms*, pages 117–151. Pontificiae Academiae Scientiarum Scripta, 1985.

[MG92] J. H. Maunsell and J. R. Gibson. Visual response latencies in striate cortex of the macaque monkey. *Journal of Neurophysiology*, 68(4):1332–1344, 1992.

[MG01] Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Comput. Vis. Image Underst.*, 81(3):231–268, 2001.

[MGA+99] JOHN H.R. MAUNSELL, GEOFFREY M. GHOSE, JOHN A. ASSAD, CARRIE J. McADAMS, CHRISTEN ELIZABETH BOUDREAU, and BRETT D. NOERAGER. Visual response latencies of magnocellular and parvocellular lgn neurons in macaque monkeys. *Visual Neuroscience*, 16(01):1–14, 1999.

[MHK06] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104:90–126, November 2006.

[MJB05] Antonio S. Micilotta, Eng Jon, and Ong Richard Bowden. Detection and tracking of humans by probabilistic body part assembly. In *Proc. of British Machine Vision Conference*, pages 429–438, 2005.

[MN84] S. P. McKee and K. Nakayama. The detection of motion in the peripheral visual field. *Vision Research*, 24:25–32, 1984.

[MN87] J H R Maunsell and W T Newsome. Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10(1):363–401, 1987.

[MN96] J. Anthony Movshon and William T. Newsome. Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys. *The Journal of Neuroscience*, 16(23):7733–7741, 1996.

[MNCG01] B. McCane, K. Novins, D. Crannitch, and B. Galvin. On benchmarking optical flow. *Comput. Vis. Image Underst.*, 84(1):126–143, 2001.

[MRS92] G. Mather, K. Radford, and West S. Low-level visual processing of biological motion. *Proc. Roy. Soc. Lond. Series B*, 249:149–155, 1992.

[MTHC03] Ivana Mikic, Mohan Trivedi, Edward Hunter, and Pamela Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53:199–223, 2003. 10.1023/A:1023012723347.

[Muy55] Eadweard Muybridge. *The human figure in motion.* New York (N.Y.) : Dover Publications, 1955.

[MVB94] A. B. Metha, A. J. Vingrys, and D. R. Badcock. Detection and discrimination of moving stimuli: the effects of color, luminance, and eccentricity. *J. Opt. Soc. Am. A*, 11(6):1697, 1994.

[MVE85] J. H. R. Maunsell and D. C. Van Essen. Functional properties of neurons in the middle temporal visual area (mt) of the macaque monkey: I. selectivity for stimulus direction, speed and orientation. *J. Neurophysiol.*, 49:1127–1147, 1985.

[NA08] Mark Nixon and Alberto S. Aguado. *Feature Extraction & Image Processing, Second Edition.* Academic Press, 2 edition, January 2008.

[NHD05] H. Nover, Anderson C. H., and G. C. DeAngelis. A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *The J. of Neurosci.*, 25(43):10049–10060, October 2005.

[OCDBM85] G. A. Orban, F. V. Calenbergh, B. De Bruyn, and H. Maes. Velocity discrimination in central and peripheral visual field. *J. Opt. Soc. Am. A*, 2(11):1836, 1985.

[OSK03] M. Okada, Nishina S., and M Kawato. The neural computation of the aperture problem: an iterative process. *Neuroreport*, 14(14):1767–1771, 2003.

[Pav00] Theo Pavlidis. 36 years on the pattern recognition front, 2000. `http://www.theopavlidis.com/technology/KSFuLecture.htm`.

[PC01] Anitha Pasupathy and Charles E. Connor. Shape representation in area v4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–2519, 2001.

[PD07] Marius V. Peelen and Paul E. Downing. The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648, 2007.

[Per92] Jacquelin Perry. *Gait Analysis: Normal and Pathological Function*. Delmar Learning, 1st edition, January 1992.

[PF03] Ralf Plänkers and Pascal Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:1182–1187, September 2003.

[PGS+07] Dumitru Petrusca, Matthew I. Grivich, Alexander Sher, Greg D. Field, Jeffrey L. Gauthier, Martin Greschner, Jonathon Shlens, E. J. Chichilnisky, and Alan M. Litke. Identification and characterization of a y-like primate retinal ganglion cell type. *The Journal of Neuroscience*, 27(41):11019–11027, October 2007.

[PM94] Tatiana Pasternak and William H. Merigan. Motion perception following lesions of the superior temporal sulcus in the monkey. *Cerebral Cortex*, 4(3):247–259, 1994.

[Pop10] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[PP03] A. Puce and D. Perrett. Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, Series B.*, 358:435–445, 2003.

[PPRF02] Daniel A. Pollen, Andrzej W. Przybyszewski, Mark A. Rubin, and Warren Foote. Spatial receptive field organization of macaque v4 neurons. *Cerebral Cortex*, 12(6):601–616, 2002.

[PS09] Jeannine Pinto and Maggie Shiffrar. The visual perception of human and animal motion in point-light displays. *Social Neuroscience*, 4(4):332–346, 2009.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1992.

[PvdH89] E Peterhans and R von der Heydt. Mechanisms of contour perception in monkey visual cortex. ii. contours bridging gaps. *The Journal of Neuroscience*, 9(5):1749–1763, 1989.

[PVVO05] H. Peuskens, J. Vanrie, K. Verfaillie, and G. A. Orban. Specificity of regions processing biological motion. *European Journal of Neuroscience*, 21(10):2864–2875, 2005.

[Ram91] Santiago Ramon y Cajal. *Cajal's Degeneration and Regeneration of the Nervous System*. Oxford University Press, Academic, 1991.

[RB00] Y. Ricquebourg and P. Bouthemy. Real-time tracking of moving persons by exploiting spatio-temporal image slices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):797 –808, aug 2000.

[RBBVdZ09] R.J. Reid, A. Brooks, D. Blair, and R. Van der Zwan. Snap! recognising implicit actions in static point-light displays. *Perception*, 38(4):613–616, 2009.

[RCKL06] Myung-Cheol Roh, Bill Christmas, Joseph Kittler, and Seong-Whan Lee. Robust player gesture spotting and recognition in low-resolution sports video. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision - ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 347–358. Springer Berlin / Heidelberg, 2006.

[RFG01] Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2:661–670, 2001.

[RI05] M. Masudur Rahman and Seiji Ishikawa. Human motion recognition using an eigenspace. *Pattern Recogn. Lett.*, 26:687–697, May 2005.

[RMR07] Timothy J. Roberts, Stephen J. McKenna, and Ian W. Ricketts. Human pose estimation using partial configurations and probabilistic regions. *Int. J. Comput. Vision*, 73:285–306, July 2007.

[Rol00] Edmund T Rolls. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205 – 218, 2000.

[RP00] Maximilian Riesenhuber and Tomaso Poggio. Models of object recognition. *Nature Neuroscience*, 3:1199–1204, 2000.

[RVR01] Simon J. Thorpe Rufin Van Rullen. Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Computation*, 13(6):1255–1283, June 2001.

[Sap06] Gilbert Saporta. *Probabilités, analyse des données et statistique*. Ed. Technip, 2006.

[SH98] Eero P. Simoncelli and David J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743 – 761, 1998.

[SH05] J. Starck and A. Hilton. Spherical matching for temporal correspondence of non-rigid surfaces. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1387 –1394 Vol. 2, oct. 2005.

[Sim99] E. P. Simoncelli. Bayesian multi-scale differential optical flow. In B. Jähne, H. Haussecker, and P. Geissler, editors, *Handbook of Computer Vision and Applications*, volume 2, chapter 14, pages 397–422. Academic Press, April 1999.

[SLC97] M Shiffrar, L Lichtey, and Sheba Heptulla Chatterjee. The perception of biological motion across apertures. *Perception & psychophysics*, 59(1):51–59, 1997.

[SLC04] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA, 2004. IEEE Computer Society.

[SMM05] Matthew A Smith, Najib J Majaj, and J Anthony Movshon. Dynamics of motion signaling by neurons in macaque area mt. *Nature Neuroscience*, 8:220–228, 2005.

[SMM10] Matthew A. Smith, Najib Majaj, and J. Anthony Movshon. Dynamics of pattern motion computation. In Guillaume S. Masson and Uwe J. Ilg, editors, *Dynamics of Visual Motion Processing: Neuronal, Behavioral and Computational Approaches*, pages 55–72. Springer, Berlin-Heidelberg, first edition, 2010.

[SP09] J. Schultz and K. S. Pilz. Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, 194(3):465–475, 04 2009.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47:7–42, April 2002.

[STEA91] RJ Snowden, S Treue, RG Erickson, and RA Andersen. The response of area mt and v1 neurons to transparent motion. *The Journal of Neuroscience*, 11(9):2768–2785, 1991.

[SvG08] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[SZ03] M. Srinivasan and S. Zhang. *Motion Cues in Insect Vision and Navigation*, pages 1193–1202. MIT Press, Cambridge, MA, 2003.

[Tay99] J. G. Taylor. Neural 'bubble' dynamics in two dimensions: foundations. *Biological Cybernetics*, 80(6):393–409, 1999.

[TDA01] Alexander Thiele, Karen R. Dobkins, and Thomas D. Albright. Neural correlates of chromatic motion perception. *Neuron*, 32(2):351–358, 2001.

[TG08] Steven M. Thurman and Emily D. Grossman. Temporal "Bubbles" reveal key features for point-light biological motion perception. *J. Vis.*, 8(3):1–11, 3 2008.

[THS+86] K Tanaka, K Hikosaka, H Saito, M Yukie, Y Fukada, and E Iwai. 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, 1986.

[TI98] Shiffrar M. Thornton IM, Pinto J. The visual perception of human locomotion. *Cogn. Neuropsychol*, 15(1):524–535, 1998.

[VF93] Goldsmith TH Varela FJ, Palacios AG. *Color vision of birds*, chapter 5, pages 77–98. MIT Press, Cambridge, 1993.

[VJ01] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[WA95] Y. Weiss and E.H. Adelson. Perceptually organized em: A framework for motion segmentation that combines information about form and motion. In *Vismod*, 1995.

[WBN85] Hans Wallach, Robert Becklen, and Donna Nitzberg. Vector analysis and process combination in motion perception. *Journal of Experimental Psychology: Human Perception and Performance*, 11(1):93 – 102, 1985.

[WN99] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174 – 192, 1999.

[WS00] Günther Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae (Wiley Series in Pure and Applied Optics)*. Wiley-Interscience, 2 edition, August 2000.

[WTNH03] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518, dec. 2003.

[WWS09] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 794 –801, june 2009.

[WXC08] Jian-Young Wu, Huang Xiaoying, and Zhang Chuan. Propagating Waves of Activity in the Neocortex: What They Are, What They Do. *Neuroscientist*, 14(5):487–502, 2008.

[XG02] Xiaohui Xie and Martin A. Giese. Nonlinear dynamics of direction-selective recurrent neural media. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65:051904, May 2002.

[YA98] Ming-Hsuan Yang and Narendra Ahuja. Gaussian mixture model for human skin color and its applications in image and video databases. In Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman, editors, *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 458–466. SPIE, 1998.