Learning mechanisms to account for the speed, selectivity and invariance of responses in the visual cortex

Timothée Masquelier

Université Toulouse III – Paul Sabatier
U.F.R. Sciences de la Vie et de la Terre

Thèse

pour obtenir le grade de
Docteur de l'Université de Toulouse
Délivré par l'Université Toulouse III – Paul Sabatier
Discipline : Neurosciences Cognitives

présentée et soutenue publiquement par

Timothée Masquelier

le 15 février 2008

Learning mechanisms to account for the speed, selectivity and invariance of responses in the visual cortex

Jury

Gustavo Deco	Rapporteur
Olivier Faugeras	Examinateur
Yves Frégnac	Rapporteur
Martin Giurfa	Examinateur
Pascal Mamassian	Examinateur
Simon Thorpe	Directeur de thèse

There may be only a few basic *learning* mechanisms underlying all this complex [brain] activity. The final explanation is likely to be in terms of the basic patterns of connections laid down in normal development, plus the *key learning algorithms* that modify those connections and other neural parameters. Thus the neocortex may well have an underlying simplicity, not at the level at which the mature brain behaves but at the way by which it arrives at that intricate behavior, based on its innate structure and guided by its rich experience of the world.

Francis Crick. The Astonishing Hypothesis. 1994. Touchstone.

Abstract

In this thesis I propose various learning mechanisms that could account for the speed, selectivity and invariance of the neuronal responses in the visual cortex. I also present the results of a relevant psychophysical experiment demonstrating that familiarity can accelerate visual processing.

In Chapter 2, I demonstrate that, in a feedforward neural model of the ventral stream, a combination of a temporal coding scheme, where the most strongly activated neurons fire first, with Spike Timing Dependent Plasticity (STDP) leads to a situation where neurons in higher order visual areas will gradually become selective to frequently occurring feature combinations. At the same time, their responses become more and more rapid. I firmly believe that such mechanisms are a key to understanding the remarkable efficiency of the primate visual system.

In Chapter 3, I present a second study, not restricted to vision, where one receiving STDP neuron integrates spikes from a continuously firing neuron population. It turns out, somewhat surprisingly, that STDP is able to find repeating spatio-temporal spike patterns and to track back through them, even when embedded in equally dense 'distractor' spike trains – a computationally difficult problem. STDP thus enables some form of temporal coding, even in the absence of an explicit time reference. Given that the mechanism exposed here is simple and cheap it is hard to believe that the brain did not evolve to use it.

One interesting prediction of the STDP models of Chapters 2 and 3 is that visual responses' latencies should decrease after repeated presentations of a same stimulus. In Chapter 4 I tested this prediction experimentally by inferring the visual processing times through behavioral measures. I used a saccadic forced-choice paradigm. The target was always the same repeating image (an interior scene), while the distractors (other interior scenes) were changing. The experiment revealed a familiarity-induced speed-up effect of about 100 ms. Most of it can be attributed to the learning of the task but a $\sim$25 ms effect corresponds to the familiarity with a given image, and is reached after a few hundred presentations. Of course this does not mean

that the STDP models of Chapters 2 and 3 are true – only that they are *plausible*.

In Chapter 5, I investigated the learning mechanisms that could account for the invariance of certain neuronal responses to some stimulus properties such as location or scale. It has been proposed that the appropriate connectivity could be learnt by passive exposure to smooth transformation sequences, and the use of a learning rule that takes into account the recent past activity of the cells: the 'trace rule'. I proposed a new variant of the trace rule that only reinforces the synapses between the most active cells, and therefore can handle cluttered environments. I applied it on V1 complex cells in the HMAX model, and demonstrated that, after exposure to *natural* videos, the learning rule was indeed able to form pools of simple cells with the same preferred orientation but with shifted receptive fields.

Taken together, these simulations suggest how the visual cortex could wire itself. While still speculative at the time of writing the models presented here all rely on widely accepted biophysical phenomena and are thus biologically plausible. The psychophysical results of Chapter 4 are compatible with the STDP models of Chapter 2 and 3.

Those last two models also demonstrate how the brain could easily make use of information encoded in the spike times. Whether these spike times contain additional information with respect to the averaged firing rates – a theory referred to as 'temporal coding' – is controversial. Given that the mechanisms proposed here are simple, efficient, and satisfy the known temporal constraints coming from the experimental literature, they provide a strong argument in favor of the use of temporal coding, at least when rapid processing is involved.

Keywords: vision, object recognition, ultra-rapid visual categorization, learning, temporal coding, spiking neurons, Spike Timing Dependent Plasticity (STDP)

Résumé

Dans cette thèse je propose plusieurs mécanismes d'apprentissage qui pourraient expliquer la rapidité, la sélectivité et l'invariance des réponses neuronales dans le cortex visuel. J'expose également les résultats d'une expérience de psychophysique pertinente, qui montrent que la familiarité peut accélérer les traitements visuels.

Au Chapitre 2, je démontre que, au sein d'un model neuronal de la voie ventrale de type 'feedfoward', la combinaison d'une part d'un schéma de codage temporel dans lequel les neurones les plus stimulés déchargent en premier, et d'autre part de la Spike Timing Dependent Plasticity (STDP), amène à une situation dans laquelle les neurones des aires de haut niveau deviennent graduellement sélectifs à des combinaisons fréquentes de primitives visuelles. En outre, les réponses de ces neurones deviennent de plus en plus rapides. Je crois fermement que de tels mécanismes sont à la base de la remarquable efficacité du système visuel du primate.

Au Chapitre 3 je présente une autre étude, non spécifique à la vision, dans laquelle un unique neurone reçoit des potentiels d'action (ou 'spikes') provenant d'une population d'afférents qui déchargent continuellement. Il s'avère, étonnamment, que la STDP permet de détecter puis de remonter des patterns de spikes spatio-temporels même s'ils sont insérés dans des trains de spikes 'distracteurs' de même densité – un problème computationnellement complexe. La STDP permet donc l'utilisation d'un codage temporel, même en l'absence d'une date de référence explicite. Etant donné que le mécanisme présenté ici est simple et peu coûteux, il est difficile de croire que le cerveau n'a pas évolué pour l'utiliser.

Une prédiction intéressante des modèles STDP des Chapitres 2 et 3 est que les latences des réponses visuelles devraient diminuer après présentations répétées d'un même stimulus. Au Chapitre 4 j'ai testé expérimentalement cette prédiction, en inférant les temps de traitement visuels à partir de mesures comportementales. J'ai utilisé un paradigme de choix forcé saccadique, avec comme cible toujours la même image répétée (une scène d'intérieur), alors que les distracteurs (également des scènes d'intérieur) changeaient. Les

résultats mettent en évidence une accélération des temps de traitement de l'ordre de 100 ms. La majeur partie de cet effet est imputable à l'apprentissage de la tâche, mais environ 25 ms correspondent a de la familiarité avec une image donnée. Ces 25 ms sont gagnées au bout de quelques centaines de présentations. Bien sûr cela ne veut pas dire que les modèles STDP des Chapitres 2 et 3 sont vrais – seulement qu'ils sont plausibles.

Au Chapitre 5 j'ai recherché les mécanismes d'apprentissage qui pourraient expliquer l'invariance de certaines réponses neuronales à certaines propriétés du stimulus visuel comme la position ou la taille. Il a été proposé que la connectivité appropriée pourrait être apprise à partir d'exposition passive à des séquences de transformations continues, et d'une règle d'apprentissage qui prend en compte l'activité de la cellule moyennée sur un passé récent : la 'trace rule'. Je propose une nouvelle variante de cette 'trace rule' qui renforce uniquement les synapses entre les cellules les plus actives, ce qui lui permet de fonctionner dans des environnements chargés. Je l'ai appliquée sur les cellules complexes de V1 dans le modèle HMAX, et on voit que, après exposition à des vidéos naturelles, la loi d'apprentissage forme des ensemble de cellules simples dont l'orientation préférée est la même, mais dont les champs récepteurs sont décalés.

Les simulations présentées ici suggèrent comment le cortex visuel pourrait s'auto-organiser. Même s'ils sont spéculatifs aujourd'hui, les modèles proposés s'appuient tous sur des mécanismes biophysiques communément admis – ils sont donc biologiquement plausibles. Les résultats de psychophysique du Chapitre 4 sont compatibles avec les modèles STDP des Chapitres 2 et 3.

Ces deux derniers modèles démontrent aussi comment le cerveau pourrait facilement tirer profit de l'information contenue dans les dates de spikes. Si ces dates contiennent d'avantage d'information par rapport au taux de décharge moyen – la théorie dite du 'codage temporel' – est controversé. Etant donné que les mécanismes proposés ici sont à la fois simples, efficaces, et satisfont les contraintes temporelles provenant de la littérature expérimentale, ils constituent un argument fort en faveur de l'utilisation de codage temporel, du moins dans les traitements rapides.

Mots-clefs : vision, reconnaissance d'objets, catégorisation visuelle ultra-rapide, apprentissage, codage temporel, neurones impulsionnels, Spike Timing Dependent Plasticity (STDP)

Acknowledgments

I would like to acknowledge first my advisor Dr. Simon Thorpe (DR1 CNRS, France) for the quality of his supervision, his permanent enthusiasm, his creative ideas and his broad scientific curiosity. His open-mindedness pushed him to take me in his team although I had very little experience nor training in neuroscience.

I would like to thank all the team of SpikeNet Technology Inc. (`http://www.spikenet-technology.com/`) for their support and for allowing me to do both applied and fundamental research. In particular the interactions with the R&D engineers Jong-Mo Allegraud and Nicolas Guilbaud were smooth and I think profitable for both parts. It was also a pleasure to work with them. I also acknowledge the Association Nationale pour la Recherche Technique (ANRT), which provided the other half of my funding through a Conventions Industrielles de Formation par la Recherche (CIFRE).

I would like to thank all the CERCO team for making the CERCO such a nice place to work, and in particular: Dr. Rufin Van Rullen (CR1 CNRS, France) for keeping an eye on my work, reading and commenting all my manuscripts before I submitted them, and giving me pointers to relevant literature; my predecessor and collaborator Dr. Rudy Guyonneau who first introduced me to Spike Timing Dependent Plasticity; Sébastien Crouzet for his precious help on psychophysical issues; Dr. Jean-Michel Hupé (CR1 CNRS) for his expertise and rigor on statistics.

Many thanks to my friends and collaborators at MIT: Thomas Serre (McGovern Institute, MIT), first – for convincing me to join the field of neuroscience, for the numerous brainstorms we had during my PhD, and for the profitable collaboration we had on invariance learning – and Prof. Tomaso Poggio (McGovern Institute, MIT) for welcoming me in his brilliant group in summer 2006 and spring 2007, and for the pertinent feedback he gave on my work.

I aknowledge the members of my thesis committee for their interest in my work: Prof. Dr. Gustavo Deco (ICREA, Spain), Dr. Olivier Faugeras (DR INRIA, France), Dr. Yves Frégnac (DR1 CNRS, France), Prof. Dr. Martin

Giurfa (UPS, France) and Dr. Pascal Mamassian (DR2 CNRS, France).

Contents

Chapter 1

Introduction

1.1 Learning is the key

Activity driven refinement of local neural networks, through synaptic plasticity and axon remodeling, is ubiquitous in developing neural systems, and is a *necessary* supplement to the genetically programmed mechanism of laying out coarse connections between brain areas (Katz and Shatz, 1996; Innocenti and Price, 2005). In some cases, this refinement must occur at a given period of development, said 'critical', otherwise the functionality of the network is irreversibly impaired. For example Hubel and Wiesel demonstrated that ocular dominance columns in the lowest neocortical visual area of cats, V1, were largely immutable after a critical period in development (Hubel and Wiesel, 1970). In congenitally blind people, the areas that would have become visual are involved in other functions, such as audition or language processing (see for example (Ofan and Zohary, 2007)). This means that an area 's functions largely emerges from experience, and are not hard-coded in the genes.

Among all the living organisms humans are probably the ones that learn the most. New born humans are far from being operational and need constant education and care for at least the first ten years of their lives. Wild children's development is severely impaired and lead to irreversible disfunctions (Benzaquén, 2006). At birth our brain volume is only 25% of its adult size (against 70% in the macaque). Most of the cerebral growth thus occurs after birth, while the organism is perceiving the outside world, and interacting with it. The acquisition of cognitive skills, and the underlying cerebral maturation and brain area specialization, thus result from complex interactions between experience and a genetically specified assembly program.

The cost of the long necessary training for humans is probably compensated by the ability to adapt to new environments and to build on knowledge

acquired by others, in particular across generations. This contrasts with more primitive organisms, which are genetically programmed to behave in a more fixed manner, but need less training.

From a computational point of view learning is arguably the key to understanding intelligence (Poggio and Smale, 2003), and has thus been studied extensively by the Artificial Intelligence community. In the context of neural networks learning can be defined as follow (Haykin, 1994):

Learning is a process by which the free parameters of a neural network are adapted through a process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place.

Among all the potential parameters the synaptic weights are probably the most important. In the cortex the number of connections from and to each neuron is in the order of a few thousands, and activity-driven synaptic regulation has been observed both *in vivo* and *in vitro*. The key of intelligence probably lies in this dense connectivity and its plasticity.

We distinguish supervised and unsupervised learning. Supervised learning requires a 'teacher', and is task-specific. For example a network can be trained to classify between faces and non-faces images from a set of labeled examples. The network is then able to *generalize* to new data to a certain extent, that is to label previously unseen images. This capacity to generalize, beyond the memory of specific examples, is critical (Poggio and Bizzi, 2004). Vapnik and Chervonenkis showed that there is an optimal VC dimension for the network (Vapnik and Chervonenkis, 1971) (the VC dimension is roughly the capacity of the network to fit any set of training data). If it is too small the network is not flexible enough to learn the training examples, let alone to generalize. If it is too big the network behaves like a look-up table: it does learn the specific training associations but does not generalize well to new data, unless a huge amount of training data is available. Humans, by being able to learn a new visual category from just a few examples, clearly outperform any machine-learning algorithm today. How we generalize so well remains a mystery.

In unsupervised (or self-organized) learning there is no external teacher to oversee the learning process. However, providing the world is not random, the network can tune itself to its statistical regularities. It can develop the ability to form internal representations for encoding features of the input and thereby to create new classes automatically (Becker, 1991). This type of learning is task-independent: it only depends on the world's statistics. Unsupervised learning presumably dominates in the lower layers of the visual

system.

1.2 Object recognition in the primate's visual cortex

The primate's visual cortex processes the information coming from the retina through the Lateral Geniculate Nucleus (LGN). It is made of several areas that are roughly hierarchically organized (Felleman and Van Essen, 1991). As can be seen on Fig. 1.1, it is generally assumed that the processing can be divided in two pathways: the so-called ventral and dorsal streams (Mishkin et al., 1983; DeYoe and Essen, 1988). The first one, also called the 'what' pathway, is primarily involved in object recognition (independently of the object location), whereas the second one, also called the 'where' pathway is mostly involved in spatial vision, object localization, and control of action (Ungerleider and Haxby, 1994). From now on I am going to focus on the ventral stream, which consists in a chain of neurally interconnected areas, including the primary visual cortex V1, and the extrastriate visual areas V2, V4 and IT.

Beyond IT, the Pre-Frontal Cortex (PFC) is thought to be involved in linking perception to memory and action. It is probably there that the categorization take place, essentially from the output of IT, using task specific circuits (Freedman et al., 2001).

1.2.1 Selectivity & invariance in the ventral stream

Robust object recognition requires both *selectivity* – so that an object (or object class) A is not confused with an object (class) B – and *invariance* – so that the object (class) A is recognized whatever its position, scale and whatever the viewpoint and lighting conditions, and eventually despite non-rigid transformations (for example facial expressions) and, for categorization, variations of shape within a class. Computer vision scientists know well how difficult this problem is. For example two face portraits of two individuals A and B are usually much more similar, in terms of low level image features, than a face and a profile portrait of A. Yet our visual system robustly extracts the identity the people in our visual fields, outperforming any computer vision system. How do we do that?

Over the last decades, a number of physiological studies in non-human primates have established several basic facts about the cortical mechanisms of recognition in the ventral stream. The accumulated evidence points towards two key features (see Fig. 1.2): from V1 to IT, there is a parallel increase in:

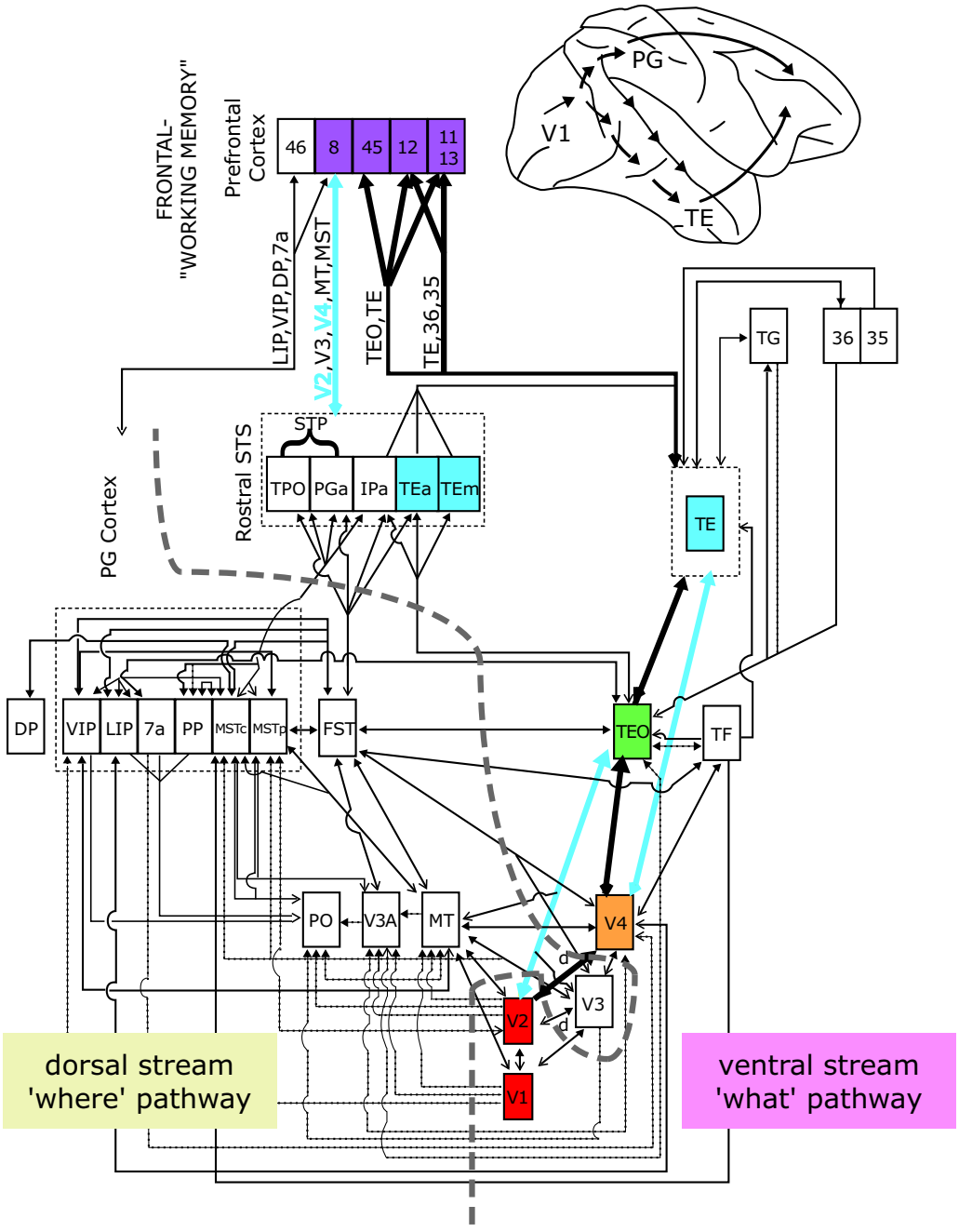


Figure 1.1: Ventral and dorsal streams of the visual cortex. Modified from Ungerleider & Van Essen (Gross, 1998). Courtesy of Thomas Serre.

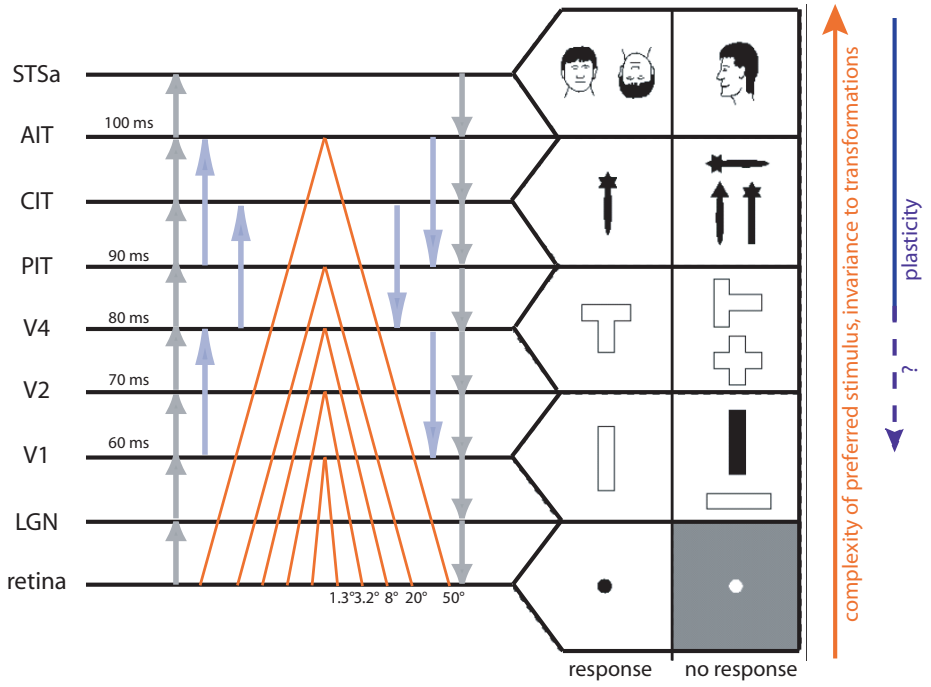


Figure 1.2: Increasing selectivity, RF sizes, and response latencies along the ventral stream. Modified it from (Oram and Perrett, 1994; Serre et al., 2007). Typical latencies are from (Thorpe and Fabre-Thorpe, 2001). RF sizes are from (Wallis and Rolls, 1997)

1. the *complexity* of the optimal stimuli for the neurons (Perrett and Oram, 1993; Desimone, 1991; Kobatake and Tanaka, 1994). That is neurons respond selectively to objects that are more and more complex. To be precise, V1 neurons' preferred stimuli are oriented bars (Hubel and Wiesel, 1959, 1968). In V2 many neurons are also orientation selective (Hubel and Wiesel, 1965, 1970) but some encode combinations of orientations such as angles (Boynton and Hegdé, 2004; Anzai et al., 2007). Further along in the hierarchy, neurons in V4 respond to features of intermediate complexity (Kobatake et al., 1998), such as Cartesian and non-Cartesian gratings (Gallant et al., 1996) or combination of boundary conformations (Pasupathy and Connor, 1999, 2001, 2002). Beyond V4, in the Infero-Temporal cortex (IT), and particularly in its anterior part (AIT) neurons are tuned to complex stimuli, for example faces, hands and other body parts (Gross, 1972; Bruce et al., 1981; Perrett et al., 1982; Rolls, 1984; Perrett et al., 1984; Baylis et al., 1985; Perrett et al., 1987; Yamane et al., 1988; Hasselmo et al., 1989; Perrett et al., 1991, 1992; Hietanen et al., 1992; Souza et al., 2005), see (Logothetis and Sheinberg, 1996) for a review.

2. *invariance* of the responses to position and scale (Hubel and Wiesel, 1968; Perrett and Oram, 1993; Logothetis et al., 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996; Riesenhuber and Poggio, 1999), and finally view point (Logothetis et al., 1995). This also means the size of the Receptive Fields (RF) increases until IT (Perrett and Oram, 1993; Tanaka, 1996). Understanding how these invariances are obtained – while neurons remain selective to their preferred stimuli – is a major challenge for visual neuroscientists. In V1, Hubel and Wiesel identified two kinds of cells that differ in their functional properties: the simple and the complex cells (Hubel and Wiesel, 1968). Both are orientation selective, but the complex cells' responses are more invariant to the phase and/or position of the stimuli. To account for this invariance, Hubel and Wiesel (1962) proposed that the complex cells could pool their inputs from a group of simple cells tuned to the same orientation, but with shifted receptive fields. A number of models have been built on this proposal, extending the scheme to the whole hierarchy (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Masquelier and Thorpe, 2007; Serre et al., 2007). How the appropriate connectivity could be learnt remains largely unknown and is the subject of Chapter 5. Lastly, one of the most striking aspect of the the shift and scale invariance observed in higher area such as IT is that it does not

seem to require exhaustive previous experience (Logothetis et al., 1995; Hung et al., 2005). Show a monkey an object it has never seen before at position P and scale S. This generates a set of IT responses R. Shifting the object by a few degrees ($\sim 4^\circ$ for typical stimulus size of 2°) or rescaling it by a few octaves (~ 2) will generate a new set of responses R' similar to R. Whether such invariance derives from a lifetime of previous experience with other similar objects (feature sharing) or from innate structural properties of the visual system or both, remains to be determined. In any case, the observation that the adult IT population has significant position and scale invariance for arbitrary 'novel' objects provides a strong constraint for any explanation of the computational architecture and function of the ventral visual stream (Hung et al., 2005).

1.2.2 Speed

The speed of object recognition in cortex is an extremely useful piece of information since it allows to infer what could be the underlying neural computations and to exclude some type of processing that are too time-consuming.

The visual system seems to have an 'fast recognition' mode – the initial phase of recognition before eye movements and high-level processes can play a role – which is already surprisingly accurate. This 'fast recognition' has been studied extensively in humans and monkeys by Thorpe and colleagues using ultra-rapid categorization paradigms (Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Rousselet et al., 2002; Bacon-Mace et al., 2005; Kirchner and Thorpe, 2006; Girard et al., 2007). Recently, it has been found that when two images are simultaneously flashed to the left and right of fixation, human subjects can make reliable saccades to the side where there is a target animal in as little as 120-130 ms (Kirchner and Thorpe, 2006). If we allow 20-30 ms for motor delays in the oculomotor system, this implies that the underlying visual processing can be done in 100 ms or less. The same protocol has just been used with monkeys, leading to even faster minimal reaction times of about 100 ms (Girard et al., 2007). Fig. 1.3 illustrates the time course of visual and motor processing in a go-no go task.

This psychophysical result is backed-up by electrophysiology in monkeys. The responses in IT begin 80-100 ms after onset of the visual stimulus, and are selective from the very beginning (Oram and Perrett, 1992), here to faces and heads, even in Rapid Serial Visual Presentation (RSVP) paradigms, when the preferred image is hidden in a continuous flow at rates up to 72 images per seconds (Keysers et al., 2001). More recently, recordings in IT showed that spike counts over time bins as small as 12.5 ms (which pro-

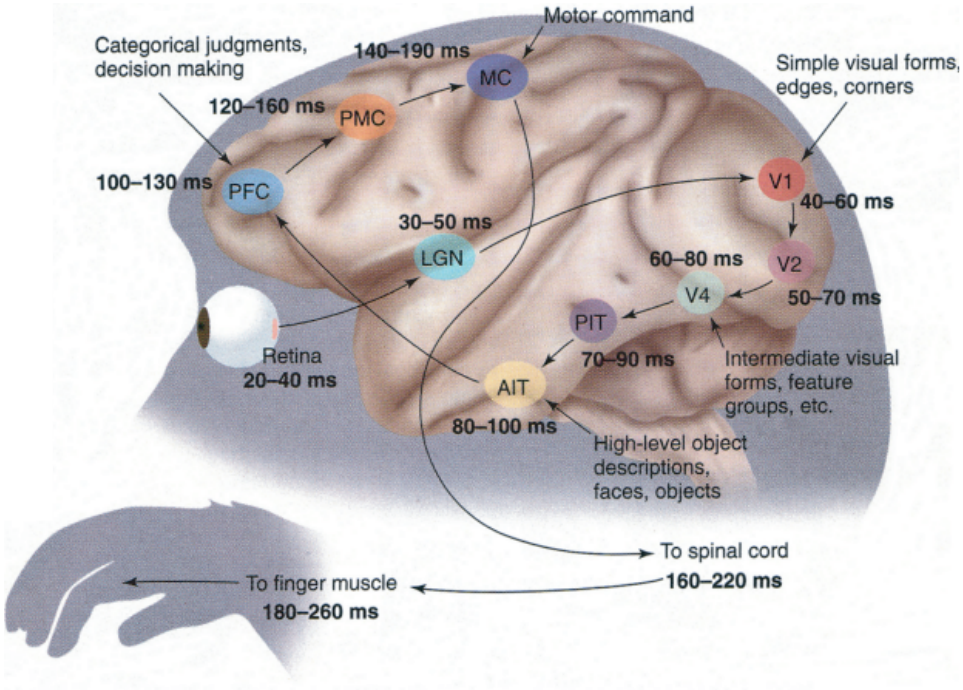


Figure 1.3: The feedforward circuits involved in a go-no go rapid visual categorization task in monkeys. At each stage two latencies are given: the first is an estimate of the earliest neuronal responses to a flashed stimulus, whereas the second provides a more typical average latency. Reproduced with permission from (Thorpe and Fabre-Thorpe, 2001)

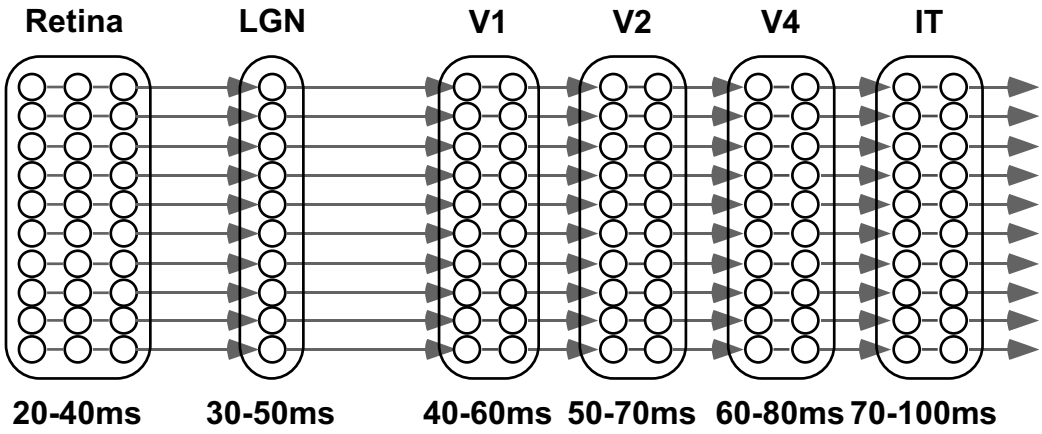


Figure 1.4: Constraints on computation time in an ultra-rapid visual categorization task (adapted from (Thorpe and Imbert, 1989)). The shortest path from the retina to IT has at least 10 neuronal layers. At each stage two latencies are given: the first is an estimate of the earliest neuronal responses to a flashed stimulus, whereas the second provides a more typical average latency. The time window available for a neuron to perform its computation is in the order of 10 ms, and will rarely contain more than one spike. Feedback is almost certainly ruled out.

duce essentially a binary vector with either ones or zeros) and only about 100 ms after stimulus onset contain remarkably accurate information about the nature of a visual stimulus (Hung et al., 2005).

These temporal constraints are extremely severe: given than about 10 neuronal layers separate IT from the retina (see Fig. 1.4), they leave about 10 ms of processing time for each neuron (Thorpe and Imbert, 1989). Since firing rates are seldom above 100 Hz in the visual system, this 10 ms window will rarely contain more than one spike. So talking about the firing rate of one neuron in this 'fast recognition' mode makes little sense, though we can talk about the firing rate of a *population* of neurons (see also 1.4.1). We can also talk about *individual spike times*. These spike times have been largely ignored by most of the neuroscientists: since Adrian recorded sensory neurons in the 1920's and reported that their mean firing rates increased with the intensity of the stimulus (Adrian, 1928) it has been commonly assumed that these rates encode most of the information. However this view is under challenge, as we will see in Section 1.3.

These severe temporal constraints have another major implication: they almost certainly rule feedback out (see Fig. 1.4), and suggests that a core hierarchical feedforward architecture may be a reasonable starting point for a theory of visual cortex aiming to explain 'fast recognition'. This hypothesis is backed up by the fact that feedforward-only models have been shown to perform very well on object recognition in natural non-segmented images (Masquelier and Thorpe, 2007; Serre et al., 2007), sometimes even matching the human performance with backward masking (Serre et al., 2007).

1.3 Learning and plasticity in the visual cortex

In the developing animal, 'rewiring' experiments (see (Horng and Sur, 2006) for a recent review), which re-route inputs from one sensory modality to an area normally processing a different modality, have now established that visual experience can have a pronounced impact on the shaping of cortical networks. How plastic is the adult visual cortex is however still a matter of debates.

From the computational perspective, it is very likely that learning may occur in all stages of the visual cortex. For instance if learning a new task involves high-level object-based representations, learning is likely to occur high-up in the hierarchy, at the level of IT or PFC. Conversely, if the task to be learned involves the fine discrimination of orientations like in perceptual learning tasks, changes are more likely to occur in lower areas at the level of V1, V2 or V4 (see (Ghose, 2004) for a review). It is also very likely

that changes in higher cortical areas should occur at faster time scales than changes in lower areas.

By now there has been several reports of plasticity in all levels of the ventral stream of the visual cortex (see (Kourtzi and DiCarlo, 2006), *i.e.* both in higher areas like PFC (Rainer and Miller, 2000; Freedman et al., 2003; Pasupathy and Miller, 2005) and IT (see for instance (Logothetis et al., 1995; Rolls, 1995; Kobatake et al., 1998; Booth and Rolls, 1998; Erickson et al., 2000; Sigala and Logothetis, 2002; Baker et al., 2002; Jagadeesh et al., 2001; Freedman et al., 2006) in monkeys or the LOC in humans (Dolan et al., 1997; Gauthier et al., 1999; Kourtzi et al., 2005; Op de Beeck et al., 2006; Jiang et al., 2007). Plasticity has also been reported in intermediate areas like in V4 (Yang and Maunsell, 2004; Rainer et al., 2004) or even lower areas like V1 (Singer et al., 1982; Karni and Sagi, 1991; Yao and Dan, 2001; Schuett et al., 2001; Crist et al., 2001), although their extent and functional significance is still under debate (Schoups et al., 2001; Ghose et al., 2002; DeAngelis et al., 1995).

Supervised learning procedures to validate Hebb's covariance hypothesis *in vivo* in the visual cortex at the cellular level have also been proposed. The covariance hypothesis predicts that a cell's relative preference between two stimuli could be displaced towards one of them by pairing its presentation with imposed increased responsiveness (through iontophoresis). It was indeed shown possible to durably change some cells' RF properties in cat primary visual cortex, such as ocular dominance, orientation preference, interocular orientation disparity and ON or OFF dominance, both during the critical developmental period (Frégnac et al., 1988) and in adulthood (McLean and Palmer, 1998; Frégnac and Shulz, 1999). More recently, a similar procedure was used to validate the Spike Timing Dependent Plasticity (see Section 1.7.1) in developing rat visual cortex (Meliza and Dan, 2006) using *in vivo* whole-cell recording.

Altogether the evidence suggests that learning plays a key role in determining the wiring and the synaptic weights of visual cortex cells.

1.4 Theoretical neuroscience

1.4.1 Rate coding, temporal coding and population coding

Where neural information processing is concerned, it is usually assumed that spikes are the basic currency for transmitting information between neurons, the reason being that they can propagate over large distances. How the brain

actually uses them to encode information remains more controversial.

Spikes have little variation in amplitude and duration (about 1 ms). They are thus fully characterized by their dates (the idea that this spike dates indeed contain information is referred to as 'temporal coding'). However many electrophysiologists report that the individual dates are not always reliable. Summarizing them by counting how many spikes occurred in a given time window (*i.e.* computing a mean firing rate) usually leads to more reproducible values. Whether information is lost in this averaging operation has been the object of an on going debate for some time. (the idea that most of the information remains in the averaged rate is referred to as 'rate coding').

The answer probably depends on the size of the time window. If too big, averaged values may fail to capture some dynamical aspects of the responses. It is thus tempting to use a small window, and compute a more 'instantaneous' firing rate. However to estimate the firing rate of one neuron the time window must contain at least a few spikes. The minimal time window is thus in the order of a few typical Inter Spike Intervals (ISI). This is sometimes longer than the order of magnitude of some behavioral times, ruling out the hypothesis that the individual firing rates are indeed used in the neural computations that underlie the behavior.

Electrophysiologists can sometimes reduce this window by averaging over several runs, with carefully controlled conditions. Obviously this solution is not possible in the brain. However the same result could be obtained by averaging over a population of redundant neurons with similar selectivity. This is referred to as 'population coding' and is indeed a possibility, though costly in terms of number of neurons (Gautrais and Thorpe, 1998).

In this thesis, I explored another possibility. I assumed individual spike times were (somewhat) reliable, despite what some electrophysiologists think, and investigated how information could be encoded and decoded in those spike times.

1.4.2 Randomness, noise, and unknown sources of variability

According to Laplace, randomness is only a measure of our "ignorance of the different causes involved in the production of events." (Laplace, 1825). Throw a dice. You cannot guess the number that will come out. But theoretically, if you knew the initial conditions (speed and position of the dice) with enough precision, and used an fine enough model, you could compute it. The more chaotic the system (high sensitivity to the initial conditions), the more you

need to know the initial conditions with accuracy. Now build a machine capable to throw the dice always at the same position and speed. If the machine is accurate enough, the the number which comes out will always be the same. Thus the dice by itself is *not* a random number generator.

Whether we live in a deterministic world or not, and the implications for the notion of free will, have been the object of a raging debate for some centuries, involving both scientists and philosophers. It is out of the scope of this thesis. However there is generally a consensus on that in realistic experiences there is always a limit on how accurately controlled the conditions are, and there are usually non-controlled ones. Both can lead to unexplained variability in the results that we call 'noise' (even though the term can be misleading and I prefer the term 'unexplained variability').

In the field of neuroscience, this variability is huge. According to the semi-serious Harvard Law of Animal Behavior: "Under carefully controlled experimental circumstances, an animal will behave as it damned well pleases". Electrophysiologists also report variability in their measures. But inferring a lack of precision in the neural code from observations of variability is hazardous. In this thesis, I will argue that the variability in some recorded spike times, in particular in the visual system, could come from non-controlled variables that might also affect neuronal activation, such as attention, eye movements, mental imagery, top-down effects *etc*. This is even more true for higher order neurons, because they do not only receive input from the retina, so the total input for these neurons is basically unknown. As Barlow wrote about neural responses in 1972, "their apparently erratic behavior was caused by our ignorance, not the neuron's incompetence." (Barlow, 1972).

1.4.3 Neuronal models

Computational neuroscientists have come-up with more or less detailed neuronal models. At which level the neurons should be modeled in a neural network model is always a difficult question. The neuronal model should capture all the essential mechanisms that underlie the network's functionality, and, to save time, avoid computing side effects which do not impact this functionality. This is easier said than done.

A somewhat coarse model is the firing rate model. It is an approximation of how a neuron behaves in a *steady regime*. The input spikes are then summarized by a firing rate x_j for each afferent (it thus ignores, among other things, that simultaneous presynaptic spikes are more efficient than distant ones in triggering a postsynaptic one). The output spikes are also

summarized as a firing rate y, given by:

$$y = f\left(\sum_j w_j \cdot x_j\right) \tag{1.1}$$

f is the transmission function that is increasing and usually non-linear. In particular f usually saturates above a certain value, because the output firing rate is limited by the refractory period. A popular choice for f is a sigmoid.

This model, although very simplified, is extremely popular, in particular among the Artificial Intelligence community. For example the well known perceptron or the Self-Organizing Maps (SOM) both use rate-based neuronal models.

In the firing rate model the individual spikes are not modeled. When needed, some authors sometimes generate them through a stochastic process (usually Poisson). The existence of such randomness in the true spike generation process is somewhat dubious, especially because we know that neurons stimulated directly by current injection in the absence of synaptic input give highly stereotyped and precise responses (Mainen and Sejnowski, 1995).

Another drawback of the firing rate model is the assumption of a steady regime. It thus fails to capture the transients, which are probably the most interesting aspects of neural computation, especially when rapid processing is involved.

For these reasons, in most the work presented here I have used the (Leaky) Integrate-and-Fire model, in which individual input and output spikes are modeled. A neuron is modeled as a capacitor C in parallel with a leaking resistance R driven by an input current $I(t)$. The membrane potential u is thus driven by:

$$C \cdot \frac{\mathrm{d}u}{\mathrm{d}t} = I(t) - \frac{u(t)}{R} \tag{1.2}$$

If we multiply Eq. 1.2 by R and introduce the membrane time constant $\tau_m = RC$ of the leaky integrator, it follows:

$$\tau_m \cdot \frac{\mathrm{d}u}{\mathrm{d}t} = R \cdot I(t) - u(t) \tag{1.3}$$

If the input current $I(t)$ is in fact generated by the arrival of presynaptic spikes s at several synapses indexed by j, with weights w_j, and at times $t_j^{(s)}$ it has the form:

$$I(t) = \sum_j w_j \cdot \sum_f \alpha(t - t_j^{(s)}) \tag{1.4}$$

where α is a kernel that expresses the current generated by one input spike

and that we will not detail here.

The LIF neuron also has a threshold. When it is reached, due to the nearly simultaneous arrival of several presynaptic spikes, a postynaptic spike is emitted. This is followed by a negative after potential and a refractory period, during which the membrane potential is set to a resting value. Then Eq. 1.3 holds again. Fig. 1.5 illustrates those points.

Finer biophysical neuronal models also exists such as conductance-based IF (gIF) models (Destexhe, 1997), the Hodgkin-Huxley model (Hodgkin and Huxley, 1952) or compartmental models (see (Brette et al., 2007) for a recent review on spiking neuron models). They provide a detailed description of how one single neuron behaves, but their computational cost is usually prohibitive for network applications like the ones I investigate in this thesis. The LIF model is widely accepted as a decent approximation of real neurons and I assumed it did capture all the essential mechanisms.

1.5 Evidence for temporal coding in the brain

Since Adrian recorded sensory neurons in the 1920's and reported that their mean firing rates increased with the intensity of the stimulus (Adrian, 1928) it has been commonly assumed that these rates encode most of the information processed by the brain. According to this view the spike generation is supposed to be a stochastic process, usually assumed to be Poisson. The signature of such a Poisson process is that the spike count over a time interval has a variance equal to its mean across trials (the ratio of both quantities is called the Fano factor and it thus equal to 1 for a Poisson process).

However the conventional view is under challenge. First various recent studies show that some neuronal responses are too reliable for the Poisson hypothesis to be tenable: for example Liu et al. (2001) and (Uzzell and Chichilnisky, 2004) find a Fano factor of about 0.3 in the retina and in the LGN respectively. Amarasingham et al. (2006) also proves that the Poisson hypothesis should be rejected for the first part of IT responses, from 100 ms to 300 ms after stimulus onset.

Second spike times are found reproducible in many neuronal systems (see Table 1.1), sometimes with millisecond precision. The time reference is either the onset of a stimulus, the maximum of a brain oscillation, or other spike times in case of spike patterns. These spike times are shown to encode information, sometimes complementary with respect to the information encoded in the rates: for example in monkeys Gawne et al. (1996) found that some V1 neurons encode the stimulus form in their rates and the stimulus contrast in their latency, and Kiani et al. (2005) showed that IT responses' latencies

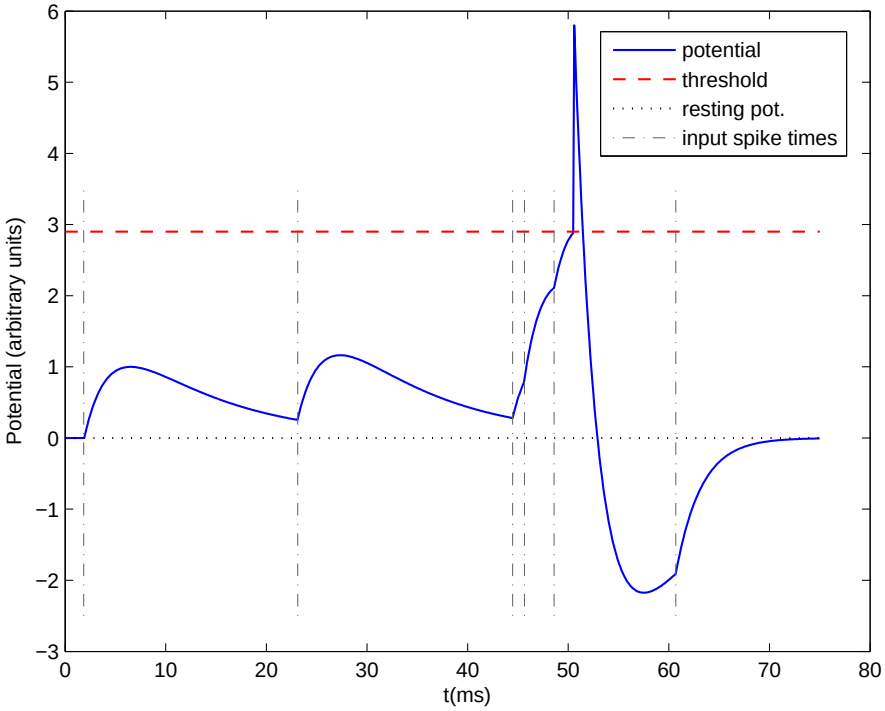


Figure 1.5: Leaky Integrate-and-Fire (LIF) neuron. Here is an illustrative example with only 6 input spikes. The graph plots the membrane potential as a function of time, and clearly demonstrates the effects of the 6 corresponding Excitatory PostSynaptic Potentials (EPSP). Because of the leak, for the threshold to be reached the input spikes need to be nearly synchronous. The LIF neuron is thus acting as a coincidence detector. When the threshold is reached, a postsynaptic spike is fired. This is followed by a refractory period of 1 ms and a negative spike-afterpotential.

were shorter for human faces than for animal faces, although both resulted in the same response magnitude.

In spontaneous activity, several second long firing sequences have also been found reproducible in slices of mouse primary visual cortex or in intact cat primary visual cortex in vivo (Ikegaya et al., 2004). Such long sequences, called 'cortical songs', could be generated by synfire chains (Abeles, 2004), that is series of pools of neurons connected in a feedforward manner. Note that the relevance of such long cortical songs in vivo is controversial because they could emerge by chance (McLelland and Paulsen, 2007; Mokeichev et al., 2007). However the spikes of the *first 100 ms* after the onset of an active period ('UP state') occur with up to millisecond precision (Luczak et al., 2007).

Other authors report a higher variability in the spike times. But first the variability could also come from the use of an inappropriate time reference. For example the stimulus onset is often used, while using the population activation onset (*i.e.* measuring the relative neuron's latency with respect to its neighbors) sometimes lead to more reproducible and informative values Chase and Young (2007). This of course requires simultaneous multi-units recording. When oscillations are present, using their maximums as a time reference (*i.e.* measuring the phase of the spikes) may lead to more reproducible values than the absolute latencies. For example Fries et al. (2001) recorded neurons in cat primary visual cortex and showed that their absolute latencies could be prolonged or shortened from one trial to another (depending on when the stimulus was presented with respect to the phase of a LFP gamma-oscillation) but their phase with respect to the gamma-cycle reference frame remained roughly constant (Fries et al., 2007).

Second, as said above in Section 1.4.2, such variability, could come from non-controlled variables.

1.6 Models of object recognition in cortex

1.6.1 Feedforward and feedback

The vast majority of models of object recognition in cortex today are feedforward only, and ignore back-projections (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Ullman et al., 2002; Masquelier and Thorpe, 2007; Serre et al., 2007). This is somewhat surprising as the circuitry of the cortex involves a massive amount of backprojections that convey information from higher areas back to the lower areas (Felleman and Van Essen,

Table 1.1: Evidence for temporal coding in the brain (adapted and updated from (VanRullen et al., 2005))

System/Preparation	Recording site	Coding	Information	Reference signal	Refs
Somatosensory					
Human	Peripheral nerve fibers	Time-to-first-spike	Direction of force, surface shape	Stimulus onset	Johansson and Birznieks (2004)
Rat	Barrel cortex	Time-to-first-spike (inhibition barrage)	Stimulus location	Stimulus onset	Petersen et al. (2001); Swadlow and Gusev (2002)
Rat	Prim. somatosens. cortex	Time-to-first-spike	Stimulus location	Stimulus onset	Foffani et al. (2004)
Olfactory					
Locust	Mushroom body	Sparse and/or binary, phase-locked (inhibition barrage)	Odor identity	20-30 Hz oscillation	Perez-Orive et al. (2002); Cassenaer and Laurent (2007)
Auditory					
Cat	Inferior colliculus	Time-to-first spike	Sound localization	Population onset	Chase and Young (2007)
Marmoset	Auditory cortex	Spike-time	Auditory event (when)	Stimulus transients	Lu et al. (2001)
Cat	Auditory cortex	Time-to-first spike	Peak pressure	Stimulus onset	Heil (1997)
Marmoset	Auditory cortex	Relative spike time	Auditory features (what)	Other spikes	deCharms and Merzenich (1996)
Rat	Auditory cortex	Time-to-first spike (binary)	Tone frequency	Stimulus onset	DeWeese et al. (2003)
Visual					
Fly	H1	Precise timing ($\sim$1 ms)	Visual event (when)	Stimulus onset & transient	de Ruyter van Steveninck et al. (1997)
Salamander	Retina	Precise timing ($\sim$3 ms)	Visual event (when)	Stimulus onset & transient	Berry and Meister (1998)
Macaque	Retina	Precise timing ($\sim$1 ms)	Visual event (when)	Stimulus onset	Uzzell and Chichilnisky (2004)
Cat	LGN	Precise timing ($<$1 ms and 1-2 ms)	Visual feature (what): luminance	Stimulus onset	Reinagel and Reid (2000); Liu et al. (2001)
Cat	LGN	Spike patterns	Visual feature (what): luminance	Other spikes	Reinagel and Reid (2002); Fellous et al. (2004)
Macaque	V1	Response latency	Visual features (what): contrast, orientation	Stimulus onset	Celebrini et al. (1993); Gawne et al. (1996)
Cat	V1	Phase and/or latency shift	Line orientation	Other spikes and/or LFP gamma oscillations	König (1995); Fries et al. (2001)
Macaque	V1	Bursts	Line orientation	Microsaccades	Martinez-Conde et al. (2000)
Macaque	V4	Phase	Identity of the stimulus to remember	LFP theta oscillation	Lee et al. (2005)
Macaque	IT	Response latency	Human vs. animal face	Stimulus onset	Kiani et al. (2005)
Macaque	MT	Precise timing (a few milliseconds)	Visual event (when)	Stimulus onset & transient	Bair and Koch (1996)
Macaque	MT	Spike patterns	Motion	Other spikes	Buracas et al. (1998); Fellous et al. (2004)
Hippocampus					
Rat	CA1 and CA3	Phase	Place	Theta oscillation	O'Keefe and Recce (1993); Mehta et al. (2002)
Premotor/prefrontal					
Macaque	Premotor/prefrontal	Spike patterns	Behavioral response	Other spikes	Prut et al. (1998)

1991). The anatomy has been extensively studied in the visual system where it is clear that feedforward connections constitute only a small fraction of the total connectivity (Douglas and Martin, 2004). For example about as many neurons project from V2 to V1 as from V1 to V2.

The main justification for these feedforward-only models is that the visual system seems to have an 'fast recognition' mode in which feedback is probably largely inactive (see Section 1.2.2). It is this mode, and only this mode that the feedforward models attempt to simulate. In this thesis, I will focus on feedforward-only models.

However, feedback and top-down mechanisms, particularly those that handle attentional effects have also been modeled. Deco and colleagues looked at top-down attention (Deco and Zihl, 2001; Deco and Lee, 2002; Rolls and Deco, 2002; Deco and Rolls, 2004, 2005). They use mean-field neurodynamical approaches in which attention is modeled as a top-down input that bias the competition between neurons of a same area. Some authors also studied bottom-up (image-based) attention which postulates that the most salient features are attended first, see for example (Tsotsos et al., 1995; Itti and Koch, 2000).

1.6.2 Static, single spike wave and mean field approximations

As for single neurons (see Section 1.4.3), the question of at which level a network should be modeled is tricky. A proper way to answer it would be to model it at the 'finest possible level', and investigate a posteriori how legitimate is a given approximation. Unfortunately it is not always possible to define such finest possible level. Furthermore this approach is often too computationally expensive. Modelers thus attempt to justify approximations a priori.

As far as the visual system modeling is concerned, three simplifications are usual, and can be used independently. The first one is the assumption of steady neuronal activities, meaning time can be removed from the equations, and it is very common (for example made by (Fukushima, 1980; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Serre et al., 2007)) Firing rates can thus be defined at the neuronal level, and rate-based neuronal models can be used (see 1.4.3). However the assumption of a steady regime is dubious. As said before there is psychophysical and electrophysiological evidence showing that high level recognition can be done in 100 ms or less in humans (see 1.2.2). This means a steady regime in terms of firing rates has not enough time to settle, at least

at the neuronal level. Models based on firing rates may thus fail to capture some key transient mechanisms of this 'fast recognition' process. This is the reason why I did not use them, except for the work on invariance learning of Chapter 5. Furthermore, static models are inherently unable to deal with dynamical stimuli, such as videos or RSVP.

The second approximation is completely different: it consists in limiting the simulation to the first spike emitted by each neuron after onset of the visual stimulus, like in most of the studies done by Thorpe and colleagues (VanRullen et al., 1998; Perrinet et al., 2001; Delorme et al., 2001; VanRullen and Thorpe, 2001, 2002; Perrinet et al., 2004b; Masquelier and Thorpe, 2007). The justification for it is that, as said in section 1.2.2, the time window available for each neuron to perform the computation which underlie 'fast recognition' is in the order of 10 ms and will rarely contain more than one spike. This approximation enormously simplifies the computations: irrespective of the actual anatomical connectivity, a network in which each neuron only ever fires one spike is by definition a pure feed-forward network, because a neuron activity cannot influence itself through any loop. Activity is thus modeled as a single spike volley (also called spike wave) that propagates across the layers of the network. Between two successive volleys all the membrane potential are reset to their resting values. This approach also means that we do not have to worry about the effects of refractory periods and synaptic dynamics such as depression due to depletion *etc.* Because of the discrete processing though, these models cannot deal with dynamical stimuli.

The third approximation is the mean field approach, in which neurons are not considered individually, but modeled in population of neurons with similar characteristics and connectivity (Deco and Zihl, 2001; Deco and Lee, 2002; Rolls and Deco, 2002; Deco and Rolls, 2004, 2005). Individual spike times are lost, but a population firing rate can be defined over small time windows. The approach is thus dynamical and can deal with dynamical stimuli. The main drawback of the method is that individual spike times are lost, which excludes the possibility that they could be informative, and spike timing-dependent phenomenon such as STDP (see Section 1.7.1) cannot be modeled.

1.6.3 Weight-sharing

Most of the bio-inspired hierarchical networks use restricted receptive fields and weight-sharing, *i.e.* each cell and its connectivity is duplicated all positions and scales (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Masquelier and Thorpe, 2007; Serre

et al., 2007) Networks using these techniques are called *convolutional* networks.

In a learning network weight-sharing allows shift-invariances to be built by structure (and not by training). It reduces the number of free parameters (and therefore the VC dimension (Vapnik and Chervonenkis, 1971)) of the network by incorporating prior information into the network design: responses should be scale and shift invariant. This greatly reduces the number of training examples needed. Note that this technique of weight sharing could be applied to other transformations than shifting and scaling, for instance rotation and symmetry.

However, it is difficult to believe that the brain could really use weight sharing. Indeed learning is problematic with such a scheme since, as noted by Földiák (Földiák, 1991), updating the weights of all the simple units connected to the same complex unit is a non-local operation. We will see in Chapter 5 how approximative weight-sharing could be implemented in the brain.

1.7 Spike Timing Dependent Plasticity (STDP)

1.7.1 Experimental evidence

Experimental studies have observed Long Term synaptic Potentiation (LTP) when a presynaptic neuron fires shortly before a postsynaptic neuron, and Long Term Depression (LTD) when the presynaptic neuron fires shortly after, a phenomenon known as Spike Timing Dependant Plasticity (STDP). The amount of modification depends on the delay between the two events: maximal when pre- and post-synaptic spikes are close together, the effects gradually decrease and disappear with intervals in excess of a few tens of milliseconds (Bi and Poo, 1998; Zhang et al., 1998; Feldman, 2000). An exponential update rule fits well the synaptic modifications observed experimentally (Bi and Poo, 2001) (see Fig. 1.6).

STDP is now a widely accepted physiological mechanism of activity-driven synaptic regulation. It has been observed extensively *in vitro* (Markram et al., 1997; Bi and Poo, 1998; Zhang et al., 1998; Feldman, 2000), and more recently *in vivo* in Xenopus's visual system (Vislay-Meltzer et al., 2006; Mu and Poo, 2006), in the locust's mushroom body (Cassenaer and Laurent, 2007), and in the rat's visual (Meliza and Dan, 2006) and barrel (Jacob et al., 2007) cortex. Very recently, it has also been shown that cortical reorganization in cat primary visual cortex is in accordance with STDP (Young et al., 2007).

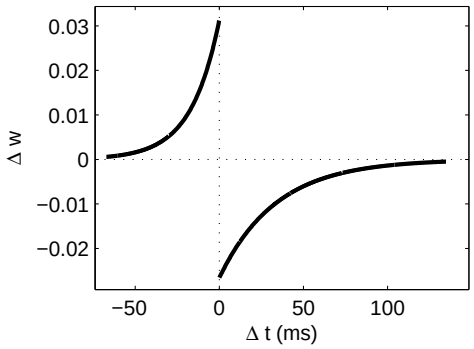


Figure 1.6: The STDP modification function. The additive synaptic weight updates as a function of the difference between the presynaptic spike time and the postsynaptic one is plotted. An exponential update rule fits well the synaptic modifications observed experimentally (Bi and Poo, 2001). The left part corresponds to Long Term Potentiation (LTP) and the right part to Long Term Depression (LTD).

Note that STDP is in agreement with Hebb's postulate because it reinforces the connections with the presynaptic neurons that fired slightly before the postsynaptic neuron, which are those which 'took part in firing it'. As a result, it *reinforces causality links*: if an input I causes the neuron N to fire, next time I occurs N is even more likely to fire, and it is also more likely to fire *earlier* with respect to the beginning of I. As we will see later these two effects of STDP are crucial.

1.7.2 Previous modeling work

STDP has received considerable interest from the modeling community over the last decade. Here I review relevant previous computational studies.

In an influential paper Song et al. (2000) demonstrated the *competitive* nature of STDP: synapses compete for the control of the postsynaptic spikes. This competition stabilizes the synaptic weights: because not all the synapses can 'win' (*i.e.* be potentiated) the sum of the synaptic weights is naturally bounded, without the need for additional normalization mechanism. Furthermore, when the system is repeatedly presented with similar spike patterns, the winning synapses are those through which the earliest spikes arrive (on average). The ultimate effect of this synaptic modification is to make the the postsynaptic neuron respond more quickly.

Gerstner and Kistler (2002) reproduced those main results and also demon-

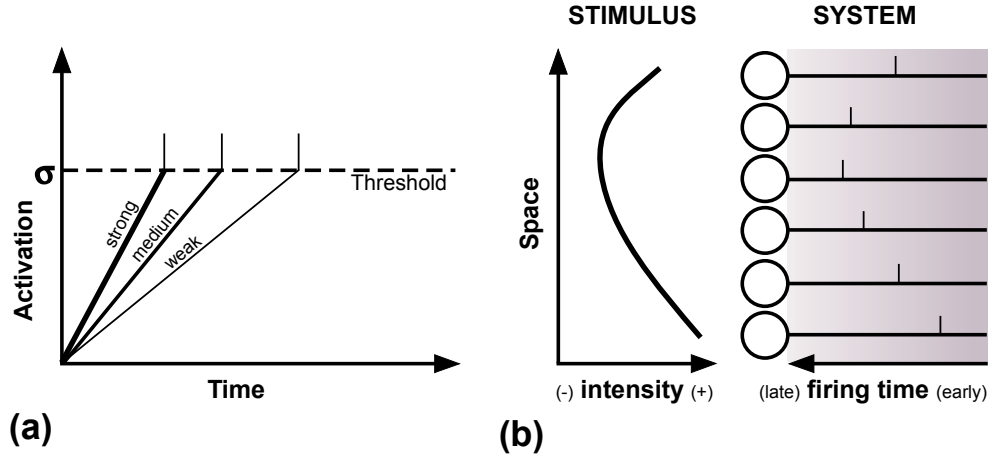


Figure 1.7: Intensity-to-latency conversion. (a) For a single neuron, the weaker the stimulus, the longer the time-to-first-spike. (b) When presented to a population of neurons, the stimulus evokes a spike wave, within which asynchrony encodes the information (reproduced with permission from Guyonneau et al. (2004))

strated that STDP increased the postynaptic spike time precision by selecting inputs with low time jitter.

Guyonneau et al. (2005) tested the robustness of Song et al. (2000)'s results in more challenging conditions, including spontaneous activity or jitter. Furthermore they also demonstrated that neither firing rate or even synchrony are relevant in the STDP selection process: only the latency matters.

STDP was also applied in visual cortex models with asynchronous spike propagation. Those models assumed one spike per neuron only (see Section 1.6.2), and an intensity-to-latency conversion in the first layer (see Fig. 1.7).

Delorme et al. (2001) and Guyonneau (2006) both showed that V1 simple cells' Gabor style orientation selectivity could emerge by applying STDP on spike trains coming from LGN ON- and OFF-center cells modeled as Difference-of-Gaussian (DoG) filters.

Guyonneau et al. (2004) used Gabor filters as inputs and propagated the same image repeatedly. The earliest spikes thus corresponded to the most salient edges of the image. By concentrating weights on the corresponding synapses STDP led to an interesting visual form detector, as can be seen on Fig. 1.8. Note that this study differs from the one presented in Chapter 2 in

that it is holistic and not feature-based.

1.8 Original contributions

1.8.1 STDP-based visual feature learning

In this study I essentially put together three ideas in the literature:

1. Multi-layer hierarchical models for robust feature-based object recognition, exemplified by (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Serre et al., 2007) (but none of these models learns in a biologically plausible manner)

2. Time-to-first spike coding. In the first layers of the network the more strongly a cell is activated the earlier it fires a spike as in (VanRullen et al., 1998; VanRullen and Thorpe, 2001).

3. STDP. Neurons at later stages of the system implement STDP, which had been shown to have the effect of concentrating the synaptic weights on afferents that systematically fire early, which causes the postsynaptic spike latency to decrease (Song et al., 2000; Gerstner and Kistler, 2002; Guyonneau et al., 2005).

I demonstrated that when such a hierarchical system is repeatedly presented with natural images, the intermediate level neurons equipped with STDP naturally become selective to patterns that are reliably present in the input, while their latencies decrease, leading to both fast and informative responses. This process occurs in an entirely unsupervised way, but I then showed that these intermediate features are able to support robust categorization.

The resulting model is appealing because it has some of the properties of other hierarchical models (robust object recognition without combinatorial explosion), but can recognize objects quickly, as has been suggested in experimental literature.

This study has been published:

Masquelier T, Thorpe SJ (2007) Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol* 3(2): e31. doi:10.1371/journal.pcbi.0030031

The original paper can be found on Section A.1. Preliminary results on

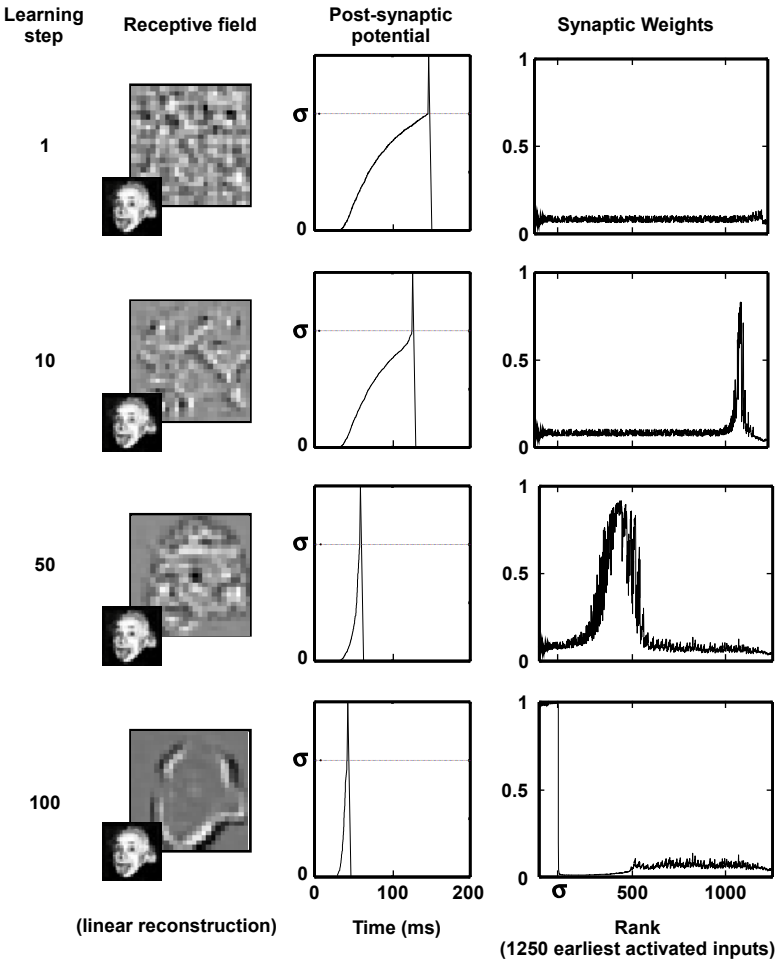


Figure 1.8: Einstein: STDP learning of a V1-filtered face. A population of V1-like cells encodes an orientation for each pixel in the image presented to the network (here, Einstein's face); each cell acts as an intensity-to-latency converter where the latency of its first spike depends on the strength of the orientation in its receptive field. Time taken to achieve recognition of the stimulus decreases (middle column) while a structured representation emerges and stabilizes (left column) that is built upon the earliest afferents of the input spike wave (right column). Note that the information concerning the evolution of the synaptic weights in the course of learning is represented twice on this figure. First in the distribution of synaptic weights on the right. It is also present in the receptive field on the left, that is linearly reconstructed based on the synaptic weights and the selectivity of the corresponding afferent neurons (here, orientation selective filters). Reproduced with permission from Guyonneau et al. (2004)

face feature learning were also presented in the conference NeuroComp 2006, Pont-à-Mousson, France (see Section B.2 for the conference paper and poster).

1.8.2 STDP-based spike pattern learning

The main limitation in the above-mentioned work on STDP-based visual feature learning – as well as in many STDP studies (Song et al., 2000; Delorme et al., 2001; Guyonneau et al., 2005; Gerstner et al., 1996) – is the assumption that the input spikes arrive in discrete volleys, each one corresponding to the presentation of one stimulus (or the maximum of a brain oscillation). The stimulus onset (or the maximum of the oscillation) is then an explicit time reference that allows defining a time-to-first spike (or latency) for each neuron. What happens in a more continuous world was unclear.

This is the reason why I started a second study, presented in Chapter 3, not restricted to vision, where one receiving STDP neuron integrates spikes from a continuously firing neuron population. It turns out, somewhat surprisingly, that STDP is able to find and track back through spatio-temporal spike patterns, even when embedded in equally dense 'distractor' spike trains – a computationally difficult problem.

STDP thus enables some form of temporal coding, even in the absence of an explicit time reference. This means that global discontinuities such as saccades or micro-saccades in vision and sniffs in olfaction, or brain oscillations in general are not necessary for STDP-based learning. Given that the mechanism exposed here is simple and cheap it is hard to believe that the brain did not evolve to use it.

This second study has also been published:

Masquelier T, Guyonneau R, Thorpe SJ (2008) Spike Timing Dependent Plasticity Finds the Start of Repeating Patterns in Continuous Spike Trains. *PLoS ONE* 3(1): e1377. doi:10.1371/journal.pone.0001377

The original paper can be found on Section A.2.

1.8.3 Visual learning experiment

One of the predictions of the two STDP models of Chapters 2 and 3 is that visual responses' latencies should decrease after repeated presentations of a same stimulus. In Chapter 4 I tested this prediction experimentally by inferring the visual processing times through behavioral measures.

In a previous study Fabre-Thorpe et al. (2001) looked for an eventual experience-induced speed-up effect, using the animal/non-animal manual

go/no-go paradigm of Thorpe et al. (1996). An extensive training with 200 animal images over a 3-week period failed to increase the speed of processing. Two reasons may explain this negative result: first we may be already experts in animal/non-animal classification, second go/no-go is usually less appropriate than saccadic forced choice to reveal subtle differences between conditions (Simon Thorpe, personal communication).

I thus used the saccadic forced-choice paradigm (Kirchner and Thorpe, 2006; Guyonneau et al., 2006; Bacon-Macé et al., 2007; Fletcher-Watson et al., 2007; Girard et al., 2007), in which one target image and one distractor image are flashed simultaneously on both sides of a fixation cross, and the participant is asked to move his eyes towards the target as fast as possible. Both accuracy (*i.e.* correct response rate) and reaction times are recorded. Here the target was always the same repeating image (an interior scene), while the distractors were changing (other interior scenes).

The experiment did reveal a familiarity-induced speed-up effect of about 100 ms. Most of it can be attributed to the learning of the task but a 25 ms effect corresponds to the familiarity with a given image, and is reached after a few hundred presentations.

1.8.4 Invariance learning

This work, presented in Chapter 5, has been done in collaboration with Thomas Serre and Tomaso Poggio, of the McGovern Institute for Brain Research, MIT. We presented preliminary results at the *VSS 07* Conference, Sarasota, FL, USA (see Section B.3 for the conference abstract and poster), and we wrote a memo:

Masquelier T, Serre T, Thorpe S and Poggio T (2007) Learning complex cell invariance from natural videos: a plausibility proof. *CBCL Paper #269 / MIT-CSAIL-TR #2007-060, Massachusetts Institute of Technology, Cambridge, MA, USA*. `http://hdl.handle.net/1721.1/39833`

We investigated the learning mechanisms that could account for the invariance of certain neuronal responses to some stimulus properties such as location or scale (see Section 1.2.1). It has been proposed that the appropriate connectivity could be learnt by passive exposure to smooth transformation sequences, and the use of a learning rule that takes into account the recent past activity of the cells: the 'trace rule' (Földiák, 1991), so as to extract slowly varying representations (Slow Feature Analysis (SFA) (Wiskott and Sejnowski, 2002) is an alternative equivalent implementation (Sprekeler et al., 2007)). However as pointed out by Spratling (2005), the trace rule by itself is

inappropriate when multiple objects are present in a scene: it cannot distinguish which input corresponds to which object, and it may end-up combining multiple objects in the same representation.

We proposed a new variant of the trace rule that only reinforces the synapses between the most active cells, and therefore can handle cluttered environments. We applied it on V1 complex cells in the HMAX model (Riesenhuber and Poggio, 1999; Serre et al., 2005a, 2007), and demonstrated that, after exposure to *natural* videos, the learning rule was indeed able to form pools of simple cells with the same preferred orientation but with shifted receptive fields.

Chapter 2

STDP-based visual feature learning

This Chapter presents an extended and updated version of the published paper (Masquelier and Thorpe, 2007):

Masquelier T, Thorpe SJ (2007) Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol* 3(2): e31. doi:10.1371/journal.pcbi.0030031

The original paper can be found on Section A.1. Preliminary results on face feature learning were also presented in the conference NeuroComp 2006, Pont-à-Mousson, France (see Section B.2 for the conference paper and poster).

2.1 Résumé

De nombreuses études expérimentales ont observé une Long Term Potentiation (LTP) quand un neurone pré-synaptique décharge peu de temps avant un neurone post-synaptique, et une Long Term Depression (LTD) quand le neurone pré-synaptique décharge peu de temps après, un phénomène connu sous le nom de Spike Timing Dependent Plasticity (STDP) (Bi and Poo, 1998; Markram et al., 1997; Zhang et al., 1998; Feldman, 2000; Vislay-Meltzer et al., 2006; Mu and Poo, 2006; Cassenaer and Laurent, 2007). Quand on présente successivement à un neurone des trains de spikes similaires en entrée on sait que STDP a pour effet de concentrer les poids synaptiques sur les afférents qui déchargent systématiquement en premier, ce qui diminue la latence du spike post-synaptique (Song et al., 2000; Guyonneau et al., 2005).

Ici, on utilise cette règle dans un réseau de neurones de type feedfor-

ward impulsionnel asynchrone qui simule la voix ventrale et l'on montre que lorsque l'on présente au réseau des images naturelles on voit progressivement émerger une sélectivité à des primitives (ou 'features') visuelles de complexité intermédiaire. Ces features, qui correspondent à des formes qui sont à la fois saillantes et présentes de manière consistante dans les images, sont très informatives et permettent une reconnaissance d'objets robuste, comme démontré sur plusieurs tâches de classification.

Le modèle est attrayant parce que, comme d'autres réseaux hiérarchiques multi-couches (du type Fukushima (1980); LeCun and Bengio (1998); Riesenhuber and Poggio (1999); Wallis and Rolls (1997); Rolls and Milward (2000); Stringer and Rolls (2000); Serre et al. (2007)), il permet une reconnaissance d'objet robuste tout en évitant une explosion combinatoire, mais aussi parce que la reconnaissance est rapide, comme suggéré par la littérature expérimentale (Oram and Perrett, 1992; Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Keysers et al., 2001; Rousselet et al., 2002; Bacon-Mace et al., 2005; Hung et al., 2005; Kirchner and Thorpe, 2006; Serre et al., 2007; Girard et al., 2007). STDP y joue un rôle clef en générant des réponses à la fois rapides et sélectives.

2.2 Abstract

Experimental studies have observed Long Term synaptic Potentiation (LTP) when a presynaptic neuron fires shortly before a postsynaptic neuron, and Long Term Depression (LTD) when the presynaptic neuron fires shortly after, a phenomenon known as Spike Timing Dependant Plasticity (STDP) (Bi and Poo, 1998; Markram et al., 1997; Zhang et al., 1998; Feldman, 2000; Vislay-Meltzer et al., 2006; Mu and Poo, 2006; Cassenaer and Laurent, 2007). When a neuron is repeatedly presented with similar inputs STDP is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early, while postsynaptic spike latencies decrease (Song et al., 2000; Guyonneau et al., 2005).

Here we use this learning rule in an asynchronous feedforward spiking neural network that mimics the ventral visual pathway, and show that when the network is presented with natural images selectivity to intermediate complexity visual features emerges. Those features, which correspond to prototypical patterns that are both salient and consistently present in the images, are highly informative and enable robust object recognition, as demonstrated on various classification tasks.

The resulting model is appealing because, like other hierarchical models Fukushima (1980); LeCun and Bengio (1998); Riesenhuber and Poggio

(1999); Wallis and Rolls (1997); Rolls and Milward (2000); Stringer and Rolls (2000); Serre et al. (2007), it is able of robust object recognition without combinatorial explosion, but it can also do it fast, as has been suggested in experimental literature (Oram and Perrett, 1992; Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Keysers et al., 2001; Rousselet et al., 2002; Bacon-Mace et al., 2005; Hung et al., 2005; Kirchner and Thorpe, 2006; Serre et al., 2007; Girard et al., 2007). By generating fast and selective responses STDP plays a key role here.

2.3 Introduction

Temporal constraints pose a major challenge to models of object recognition in cortex. There is now psychophysical and electrophysiological evidence showing that object recognition occurs in 100 ms or less in humans (see Section 1.2.2). This 'fast recognition' presumably depends on the ability of the visual system to learn to recognize familiar visual forms in an unsupervised manner. Quite how this learning occurs constitutes a major challenge for theoretical neuroscience.

Here we explored the capacity of network architectures that have three key features. First we used a feedforward multi-layer hierarchical model of the kind exemplified by (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Serre et al., 2007). These networks are known to be able of robust feature-based object recognition. Second, when stimulated with a flashed visual stimulus, the neurons in the various layers of the system fire asynchronously, with the most strongly activated neurons firing first – a mechanism that has been shown to efficiently encode image information (VanRullen and Thorpe, 2001). Third, neurons at later stages of the system implement Spike-Time Dependent Plasticity, which is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early (Song et al., 2000; Guyonneau et al., 2005).

We demonstrate that when such a hierarchical system is repeatedly presented with natural images, these intermediate level neurons will naturally become selective to patterns that are reliably present in the input, while their latencies decrease, leading to both fast and informative responses. This process occurs in an entirely unsupervised way, but we then show that these intermediate features are able to support robust categorization.

2.4 Model

2.4.1 Hierarchical architecture

Our network belongs to the family of feedforward hierarchical convolutional networks used by many other studies (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2007). To be precise its architecture is inspired from Serre, Wolf and Poggio's model of object recognition (Serre et al., 2005b), a model that itself extends HMAX (Riesenhuber and Poggio, 1999) and performs remarkably well with natural images. Like them, in an attempt to model the increasing complexity and invariance observed along the ventral pathway (see Section 1.2.1), we use a four layer hierarchy (S_1–C_1–S_2–C_2) where simple cells (S) gain their selectivity from a linear sum operation, while complex cells (C) gain invariance from a nonlinear max pooling operation (see Fig. 2.1, and Section 2.7 for a complete description of our model).

2.4.2 Temporal coding

Nevertheless our network does not only rely on static non-linearities: it uses spiking neurons and operates in the temporal domain. At each stage the time-to-first spike with respect to stimulus onset (or to be precise the rank of the first spike in the spike train as we will see later), is supposed to be the 'key variable', that is the variable which contains information and which is indeed read-out and processed by downstream neurons. When presented with an image, the first layer's S_1 cells, emulating V1 simple cells, detect edges with four preferred orientations and the more strongly a cell is activated the earlier it fires. This intensity-latency conversion (see Fig. 1.7) is in accordance with recordings in V1 showing that response latency decreases with the stimulus contrast (Albrecht et al., 2002; Gawne et al., 1996) and with the proximity between the stimulus orientation and the cell's preferred orientation (Celebrini et al., 1993). It has already been shown how such orientation selectivity can emerge in V1 by applying STDP on spike trains coming from retinal ON and OFF-center cells (Delorme et al., 2001; Guyonneau, 2006), so for simplicity we started our model from V1 orientation-selective cells. We also limit the number of spikes at this stage by introducing competition between S_1 cells through a 1-Winner-Take-All mechanism: at a given location – corresponding to one cortical column – only the spike corresponding to the best matching orientation is propagated (sparsity is thus 25% at this stage). Note that k-Winner-Take-All mechanisms are easy to implement in the temporal domain using inhibitory GABA interneurons (Thorpe, 1990).

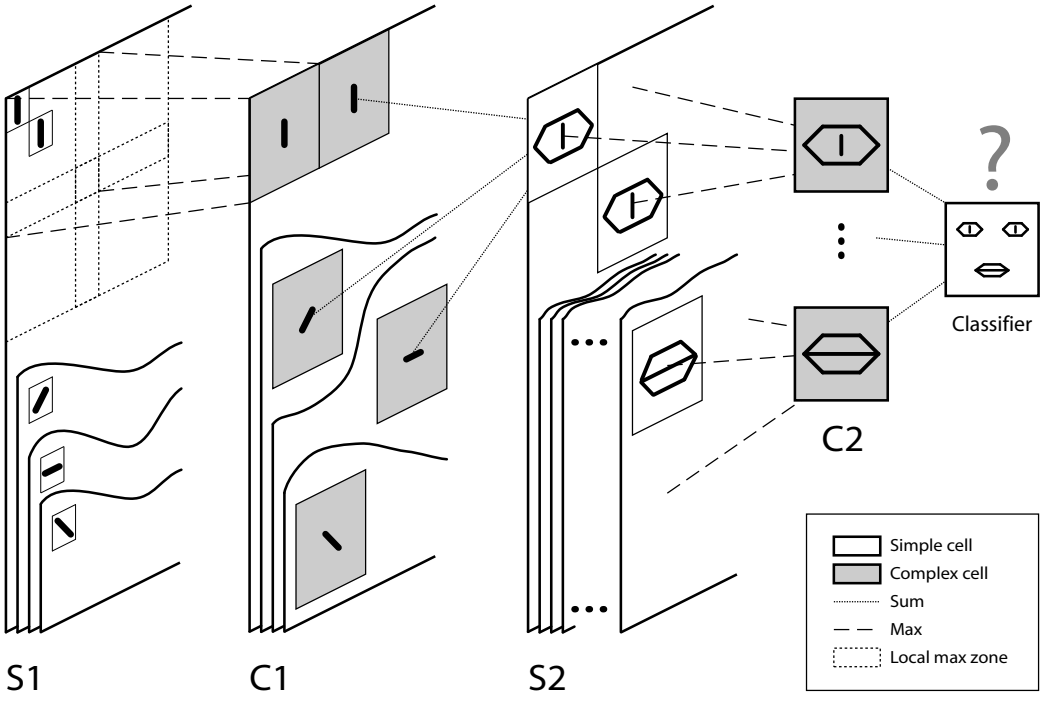


Figure 2.1: Overview of the 5 layer feedforward spiking neural network. As in HMAX (Riesenhuber and Poggio, 1999) we alternate simple cells that gain selectivity through a sum operation, and complex cells that gain shift and scale invariance through a max operation (that simply consists in propagating the first received spike). Cells are organized in retinotopic maps until the S_2 layer (included). S_1 cells detect edges. C_1 maps sub-sample S_1 maps by taking the maximum response over a square neighborhood. S_2 cells are selective to intermediate complexity visual features, defined as a combination of oriented edges (here we symbolically represented an eye detector and a mouth detector). There is one S_1–C_1–S_2 pathway for each processing scale (not represented on the figure). Then C_2 cells take the maximum response of S_2 cells over all positions and scales and are thus shift and scale invariant. Finally, a classification is done based on the C_2 cells' responses (here we symbolically represented a face / non face classifier). In the brain equivalents of S_1 cells may be in V1, of S_2 cells in V1-V2, S_2 cells in V4-PIT, C_2 cells in AIT and the final classifier in PFC. This chapter focuses on the learning of C_1 to S_2 synaptic connections through STDP.

These S_1 spikes are then propagated asynchronously through the feed-forward network of integrate-and-fire neurons. Note that within this time-to-first-spike framework, the maximum operation of complex cells simply consists in propagating the first spike emitted by a given group of afferents (Rousselet et al., 2003a). This can be done efficiently with an integrate-and-fire neuron with low threshold that has synaptic connections from all neurons in the group.

Images are processed one by one, and we limit activity to at most one spike per neuron, that is, only the initial spike wave is propagated. Before presenting a new image, every neuron's potential is reset to zero. We process various scaled versions of the input image (with the same filter size). There is one S_1-C_1-S_2 pathway for each processing scale (not represented on Fig. 2.1). This results in S_2 cells with various receptive field sizes (see Section 2.7). Then C_2 cells take the maximum response (*i.e.* first spike) of S_2 cells over all positions and scales, leading to position and scale invariant responses.

This chapter explains how STDP can set the C_1-S_2 synaptic connections, leading to intermediate complexity visual features, whose equivalent in the brain may be in V4 or IT. STDP is a learning rule that modifies the strength of a neuron's synapses as a function of the precise temporal relations between pre- and post-synaptic spikes (see Section 1.7.1). Here we used a simplified STDP rule where the weight modification does not depend on the delay between pre- and post-synaptic spikes, and the time window is supposed to cover the whole spike wave (see Section 2.7.5). We also use 0 and 1 as 'soft bounds', ensuring the synapses remain excitatory. The effects of STDP with Poisson spike trains have been studied (Song et al., 2000; VanRossum et al., 2000). Here, we demonstrate STDP's remarkable ability to detect statistical regularities in terms of earliest firing afferent patterns within visual spike trains, despite their very high dimensionality inherent to natural images.

2.4.3 STDP-based learning

Visual stimuli are presented sequentially and the resulting spike waves are propagated through to the S_2 layer, where STDP is used. We use restricted receptive fields (*i.e.* S_2 cells only integrate spikes from a s x s square neighborhood in the C_1 maps corresponding to one given processing scale) and weight sharing (*i.e.* each prototype S_2 cell is duplicated in retinotopic maps and at all scales). Starting with a random weight matrix (size = 4 x s x s) we present the first visual stimuli. Duplicated cells are all integrating the spike train and compete with each other. If no cell reaches its threshold nothing happens and we process the next image. Otherwise for each prototype the first duplicate to reach its threshold is the winner. A 1-Winner-Take-All

mechanism prevents the other duplicated cells from firing. The winner thus fires and the STDP rule is triggered. Its weight matrix is updated, and the change in weights is duplicated at all positions and scales. This allows the system to learn patterns despite of changes in position and size in the training examples. We also use local inhibition between different prototype cells: when a cell fires at a given position and scale, it prevents all other cells from firing later at the same scale and within an s/2 x s/2 square neighborhood of the firing position. This competition, only used in the learning phase, prevents all the cells from learning the same pattern. Instead, the cell population self-organizes, each cell trying to learn a distinct pattern so as to cover the whole variability of the inputs.

If the stimuli have visual features in common (which should be the case if for example they contain similar objects), the STDP process will extract them. That is, for some cells we will observe convergence of the synaptic weights (by saturation), which end up being either close to 0 or to 1. During the convergence process synapses compete for control of the timing of postsynaptic spikes (Song et al., 2000). The winning synapses are those through which the earliest spikes arrive (on average) (Song et al., 2000; Guyonneau et al., 2005), and this is true even in the presence of jitter and spontaneous activity (Guyonneau et al., 2005) (although the model presented in this chapter is fully deterministic). This 'preference' for the earliest spikes is a key point since the earliest spikes, which correspond in our framework to the most salient regions of an image, have been shown to be the most informative (VanRullen and Thorpe, 2001). During the learning the postsynaptic spike latency decreases (Song et al., 2000; Guyonneau et al., 2005; Gerstner and Kistler, 2002). After convergence, the responses become selective (in terms of latency) (Guyonneau et al., 2005) to visual features of intermediate complexity similar to the features used in earlier work (Ullman et al., 2002). Features can now be defined as clusters of afferents that are consistently among the earliest to fire. STDP detects these kinds of statistical regularities among the spike trains and creates one class of units for each distinct pattern.

2.5 Results

We evaluated our STDP-based learning algorithm on two Caltech datasets, one containing faces and the other motorbikes, and a distractor set containing backgrounds, all available at `www.vision.caltech.edu` (see Fig. 2.2 for sample pictures). Note that most of the images are not segmented. Each dataset was split into a training set used in the learning phase, and a testing set, not

Figure 2.2: Sample pictures from the Caltech datasets. The top row shows examples of faces (all unsegmented), the middle row shows examples of motorbikes (some are segmented, others are not), and the bottom row shows examples of distractors.

seen during the learning phase, but used afterwards to evaluate the performance on novel images. This standard cross-validation procedure allows the measurement of the system's ability to *generalize*, as opposed to learning the specific training examples. The splits used were the same as Fergus et al. (2003). All images were rescaled to be 300 pixels in height (preserving the aspect ratio) and converted to gray-scale values.

2.5.1 Single-class

We first applied our unsupervised STDP-based algorithm on the face and motorbike training examples (separately), presented in a random order, to build two sets of ten class specific C_2 features. Each C_2 cell has one preferred input, defined as a combination of edges (represented by C_1 cells). Note that many gray level images may lead to this combination of edges, because of the local max operation of C_1 cells, and because we lose the 'polarity' information (*i.e.* which side of the edge is darker). However we can reconstruct a representation of the set of preferred images by convolving the weight matrix with a set of kernels representing oriented bars. Since we start with random weight matrices, at the beginning of the learning process the reconstructed preferred stimuli do not make much sense. But as the cells learn, structured

representations emerge, and we are usually able to identify the nature of the cells' preferred stimuli. Fig. 2.3 and 2.4 show the reconstructions at various stages of learning for the face and motorbike datasets respectively. We stopped the learning after 10,000 presentations.

Then we turned off the STDP rule and tested these STDP-obtained features' ability to support face / non face and motorbike / non motorbike classification. This chapter focuses more on feature extraction than on sophisticated classification methods, so we first used a very simple decision rule based on the number of C_2 cells that fired with each test image, on which a threshold is applied. Such a mechanism could be easily implemented in the brain. The threshold was set to be at equilibrium point (*i.e.* when the false positive rate equals the missed rate). In Table 2.1 we report good classification results with this 'simple count' scheme in terms of area under the Receiver Operator Characteristic (ROC) and the performance rate at equilibrium point.

We also evaluated a more complicated classification scheme. C_2 cells' thresholds were supposed to be infinite, and we measured the final potentials they reached after having integrated the whole spike train generated by the image. This final potential can be seen as the number of early spikes in common between a current input and a stored prototype (this contrasts with HMAX and extensions (Riesenhuber and Poggio, 1999; Serre et al., 2005b, 2007), where an Euclidian distance or a normalized dot product is used to measure the difference between a stored prototype and a current input). Note that this potential is contrast invariant: a change in contrast will shift all the latencies, but will preserve the spike order. The final potentials reached with the training examples were used to train a Radial Basis Function (RBF) classifier (see Section 2.7.6). We chose this classifier because linear combination of Gaussian-tuned units is hypothesized to be a key mechanism for generalization in the visual system (Poggio and Bizzi, 2004). We then evaluated the RBF on the testing sets. As can be seen in Table 2.1, performance with this 'potential+RBF' scheme was better.

Using only ten STDP-learnt features we reached on those two classes a performance that is comparable to that of Serre, Wolf and Poggio's model, that itself is close to the best state-of-the-art computer vision systems (Serre et al., 2005b). However their system is more generic. Classes with more intra-class variability (for example, animals) appear to pose a problem with our approach, because a lot of training examples (say a few tens) of a given feature type are needed for the STDP process to learn it properly.

Our approach leads to the extraction of a small set (here ten) of highly informative class-specific features. This is in contrast with Serre et al's approach where many more (usually about a thousand) randomly extracted

Figure 2.3: Evolution of reconstructions for face features. Above is the number of postsynaptic spikes emitted. Starting from random preferred stimuli, cells detect statistical regularities among the input visual spike trains after a few hundred discharges, and progressively develop selectivity to those patterns. A few hundred more discharges are needed to reach a stable state. Furthermore, the population of cells self-organizes, with each cell effectively trying to learn a distinct pattern so as to cover the whole variability of the inputs.

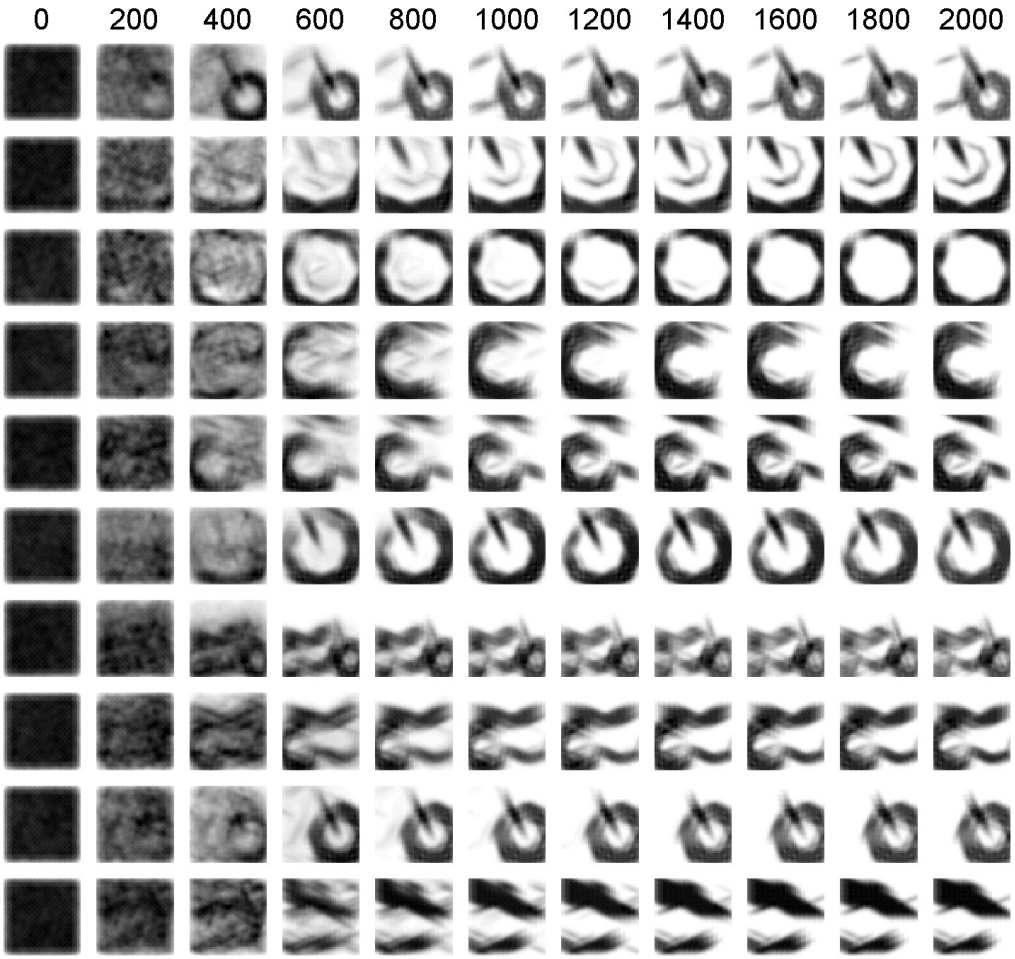


Figure 2.4: Evolution of reconstructions for motorbike features.

features are used. Their sets are more generic, and suitable for many different classes (Serre et al., 2005b). They rely on the final classifier to 'select' diagnostic features and appropriately weight them for a given classification task. Here, STDP will naturally focus on what is common to the positive training set, *i.e.* target object features. The background is generally not learned (at least not in priority), since backgrounds are almost always too different from one image to another for the STDP process to converge. Thus we directly extract diagnostic features, and we can obtain reasonably good classification results using only a threshold on the number of detected features. Furthermore as STDP performs vector quantization from multiple examples as opposed to 'one shot learning', it will not learn the noise, nor anything too specific to a given example, with the result that it will tend to learn archetypical features.

Another key point is the natural trend of the algorithm to learn salient regions, simply because they correspond to the earliest spikes, with the result that neurons whose receptive fields cover salient regions are likely to reach their threshold (and trigger the STDP rule) before neurons 'looking' at other regions. This contrasts with more classical competitive learning approaches, where input normalization helps different input patterns to be equally effective in the learning process (Rolls and Deco, 2002). Note that 'salient' means within our network 'with well defined contrasted edges', but saliency is a more generic concept of local differences, for *e.g.* in intensity, color, orientations as in Itti et al's model (Itti et al., 1998). We could use other types of S_1 cells to detect other types of saliency, and provided we apply the same intensity-latency conversion, STDP would still focus on the most salient regions. Saliency is known to drive attention (see (Treue, 2003) for a review). Our model predicts that it also drives the learning. Future experimental work will test this prediction.

2.5.2 Multi-class

Of course, in real life we are unlikely to see many examples of a given category in a row. That is why we performed a second simulation, where twenty C_2 cells were presented with the face, motorbike, and background training pictures in a random order, and the STDP rule was applied. Fig. 2.5 shows all the reconstructions for this mixed simulation after 20,000 presentations. We see that the twenty cells self-organized, some of them having developed selectivity to face features, and others to motorbike features. Interestingly, during the learning process the cells rapidly show a preference for one category. After a certain degree of selectivity has been reached, the face feature learning is not influenced by the presentation of motorbikes (and vice versa),

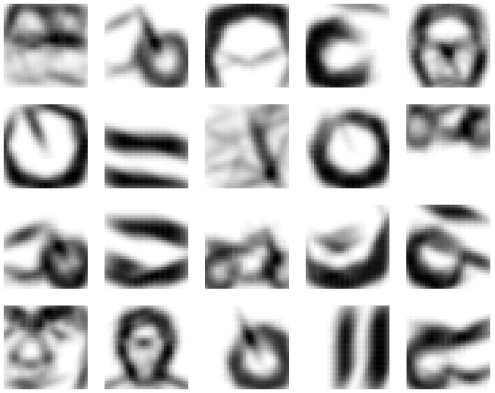


Figure 2.5: Final reconstructions for the twenty features in the mixed case. The twenty cells self-organized, some having developed selectivity to face features, and some to motorbike features.

simply because face cells will not fire (and trigger the STDP rule) on motorbikes.

Again we tested the quality of these features with a (multi-class) classification task, using an RBF network and a 'one versus all' approach (see Section 2.7.6). As before we tested two implementations: one based on a 'binary detections+RBF' and one based on 'potential+RBF'. Note that a simple detection count can not work here, as we need at least some supervised learning to know which feature (or feature combination) is diagnostic (or anti-diagnostic) of which class. Table 2.2 shows the confusion matrices obtained on the testing sets for both implementations, leading respectively to 95.0% and 97.7% of correct classifications on average. It is worth mentioning that the 'potential+RBF' system perfectly discriminates between faces and motorbikes – although both were presented in the unsupervised STDP-based learning phase.

Table 2.1: Classification results

Model	STDP (Simple Count)		STDP (Potential+RBF)		Hebbian		Serre et al. (2005b)	
	Equilibrium Pt	ROC	Equilibrium Pt	ROC	Equilibrium Pt	ROC	Equilibrium Pt	ROC
Face	96.5	99.1	99.1	100.0	96.9	99.7	98.2	99.8
Motorbikes	95.4	98.4	97.8	99.7	96.5	99.3	98	99.8

Table 2.2: Confusion matrices

Prediction with:	STDP Features (Binary)			STDP Features (Potential)			Hebbian Features		
	Face	Motorbike	Background	Face	Motorbike	Background	Face	Motorbike	Background
Actual Face	97.2	0.5	2.3	98.2	0.0	1.8	97.7	0.0	2.3
Actual Motorbike	0.0	95.3	4.8	0.0	97.5	2.5	0.3	96.3	3.5
Actual Background	3.1	4.4	92.4	0.4	2.2	97.3	4.9	3.6	91.6

2.5.3 Hebbian learning

An interesting control is to compare the STDP learning rule with a more standard hebbian rule, in this precise framework. For this purpose we converted the spike trains coming from C_1 cells into a vector of (real valued) C_1 activities X_{C_1}, supposed to correspond to firing rates (see Section 2.7.7). Each S_2 cell was not modeled anymore at the integrate-and-fire level, but was supposed to respond with a (static) firing rate Y_{S_2} given by the normalized dot product:

$$Y_{S_2} = \frac{W_{S_2} \cdot X_{C_1}}{|X_{C_1}|_2} \tag{2.1}$$

where W_{S_2} is the synaptic weight vector of the S_2 cell, and $|.|_2$ is the usual euclidian norm.

The S_2 cells still competed with each other, but the kWTA mechanisms now selected the cells with the highest firing rates (instead of the first one to fire). Only the cells whose firing rates reached a certain threshold were considered in the competition (see Section 2.7.7). The winners now triggered the following modified hebbian rule (instead of STDP):

$$\delta W_{S_2} = a \cdot Y_{S_2} \cdot (X_{C_1} - W_{S_2}) \tag{2.2}$$

where a decay term has been added in order to keep the weight vector bounded (however the rule is still local, unlike the situation with an explicit weight normalization). Note that this precaution was not needed in the STDP case, because competition between synapse naturally bounds the weight vector (Song et al., 2000). The rest of the network is strictly identical to the STDP case.

Fig. 2.6 shows the reconstruction of the preferred stimuli for the ten C_2 cells after 10,000 presentations for the face stimuli (top) and the motorbikes stimuli (bottom). Again we can usually recognize the face and motorbike parts to which the cells became selective (even though the reconstructions look fuzzier than in the STDP case, because the final weights are more graded). We also tested the ability of these hebbian-obtained features to support face / non face and motorbike / non motorbike classification once fed into a RBF, and the results are shown in Table 2.1 (last column). We also evaluated the hebbian features with the multi-class set up. Twenty cells were presented with the same mix of face, motorbike and background pictures as before. Fig. 2.7 shows the final reconstructions after 20,000 presentations, and Table 2.2 shows the confusion matrix (last columns).

The main conclusion is that the modified hebbian rule is also able to extract pertinent features for classification (although performance on these

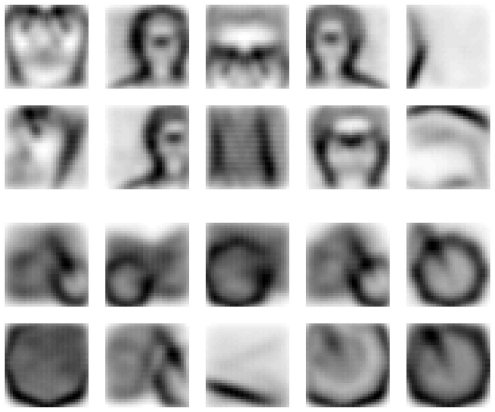


Figure 2.6: Hebbian learning. (Top) Final reconstructions for the ten face features. (Bottom) Idem for the ten motorbike features.

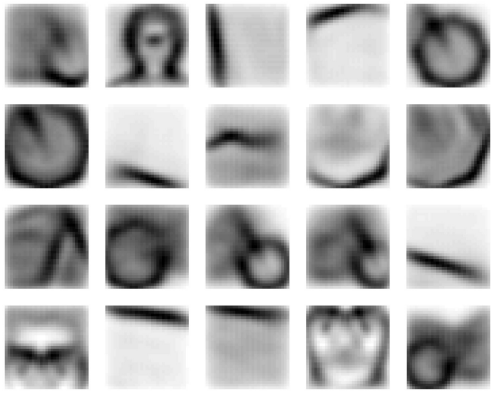


Figure 2.7: hebbian learning. Final reconstructions for the twenty features in the mixed case. As with STDP-based learning, the twenty cells self-organized, some having developed selectivity to face features, and some to motorbike features.

tests appears to be slightly worse). This is not very surprising as STDP can be seen as a hebbian rule transposed in the temporal domain, but it was worth checking it. Where STDP would detect (and create selectivity to) sets of units that are consistently among the first one to fire, the Hebbian rule detects (and creates selectivity to) sets of units that consistently have the highest firing rates. However, we believe the temporal framework is a better description of what really happens at the neuronal level, at least in ultra-rapid categorization tasks. Furthermore STDP also explains how the system becomes faster and faster with training, since the neurons learn to decode the first information available at their afferents' level (see also the Section 2.6).

2.6 Discussion

2.6.1 On learning visual features

While the ability of hierarchical feedforward networks to support classification is now reasonably well established (for *e.g.* (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000; Serre et al., 2005b, 2007)), how intermediate complexity features can be learned remains an open problem, especially with cluttered images. In the original HMAX model S_2 features were not learned but manually hard-wired (Riesenhuber and Poggio, 1999). Later versions use huge sets of random crops (say 1,000) taken from natural images, and use these crops to 'imprint' S_2 cells (Serre et al., 2005b, 2007). This approach works well and can be applied to a wide range of categorisation problems but is costly since redundancy is very high between features, and many features are irrelevant for most (if not all) the tasks. To select only pertinent features for a given task, Ullman proposed an interesting criterion based on mutual information (Ullman et al., 2002), leaving the question of possible neural implementation open. LeCun showed how visual features in a convolutional network could be learned in a supervised manner using back-propagation (LeCun and Bengio, 1998), without claiming this algorithm was biologically plausible. Although we may occasionally use supervised learning to create a set of features suitable for a particular recognition task, it seems unrealistic that we need to do that each time we learn a new class. Furthermore, a purely supervised system would be very unlikely to survive. Animals have to learn without having a teacher available. Here we took another approach: one layer with unsupervised competitive learning is used as input for a second layer with supervised learning. Note that this kind of hybrid

scheme has been found to learn much faster than a two layer backpropagation network (Rolls and Deco, 2002).

2.6.2 A bottom-up approach

Our approach is a bottom up one: instead of intuiting good image processing schemes and discussing their eventual neural correlates, we took known biological phenomena that occur at the neuronal level, namely intensity-to-latency conversion, integrate-and-fire and STDP, and saw where it could lead at a more integrated level. The role of the simulations with natural images is thus to provide a 'plausibility proof' that such mechanisms could be implemented in the brain.

2.6.3 Four simplifications

However, we have made four main simplifications. The first one was to propagate input stimuli one by one. This may correspond to what happens when an image is flashed in an ultra-rapid categorization paradigms (Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Rousselet et al., 2002; Bacon-Mace et al., 2005; Kirchner and Thorpe, 2006; Serre et al., 2007; Girard et al., 2007), but normal visual perception is an ongoing process. However, every 200 or 300 ms we typically perform a saccade. The processing of each of this discrete 'chunks' seems to be optimized for rapid execution (Uchida et al., 2006), and we suggest that much can be done with the feedforward propagation of a single spike wave. Furthermore, even when fixating, our eyes are continuously making microsaccades which could again result in repetitive waves of activation. This idea is in accordance with electrophysiological recordings showing that V1 neuron activity is correlated with microsaccades (Martinez-Conde et al., 2000). Here we assumed the successive waves did not interfere, which does not seem too unreasonable given that the neuronal time constants (integration, leak, STDP window) are in the range of a few tens of milliseconds whereas the interval between saccades and microsaccades is substantially longer. Furthermore, it is known that visual signal transmission is at least partially blocked during the saccade, presumably at the LGN level (Thiele et al., 2002). The 50-60 ms of 'pause' may be just what the cortex needs to return to a resting state, so that successive waves do not interfere. Note that this simplification allows us to use non-leaky integrate-and-fire neurons, and an infinite STDP time window. More generally, as proposed by Hopfield (Hopfield, 1995), waves could be generated by population oscillations that would fire each cell at a time in advance of the maximum of the oscillation which increases with the inputs the cell received. There is experimental

evidence for such a coding scheme in the cat primary visual cortex (König, 1995; Fries et al., 2001), see (Fries et al., 2007) for a review. This 'one-by-one processing' approximation is also legitimized by the study presented in Chapter 3, where we will see that even in a continuous regime where afferents fire continuously with a constant population rate STDP is still able to detect and learn a repeating spatio-temporal spike pattern.

The second main simplification we have made consists in using restricted receptive fields and weight sharing (see Section 1.6.3), as do most of the bio-inspired hierarchical networks (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2005b, 2007). However, it is difficult to believe that the brain could really use weight sharing since, as noted by Földiák (Földiák, 1991), updating the weights of all the simple units connected to the same complex unit is a non-local operation. Instead he suggested that at least the low level features could be learned locally and independently. Subsequently, cells with similar preferred stimulus may connect adaptively to the same complex cell, possibly by detecting correlation across time thanks to a trace rule (Földiák, 1991). Wallis, Rolls and Milward successfully implemented this sort of mechanism in a multilayered hierarchical network called Vis-Net (Wallis and Rolls, 1997; Rolls and Milward, 2000), however performance after learning objects from unsegmented natural images was poor (Stringer and Rolls, 2000). Slow Feature Analysis (Wiskott and Sejnowski, 2002), which also aims at extracting invariant representation based on the fact that they vary slowly in time, has recently been shown equivalent to Földiák's trace rule (Sprekeler et al., 2007). Future work will evaluate the use of local learning and adaptive complex pooling in our network, instead of exact weight sharing (see Chapter 5 for the type of mechanisms we would like to implement). Learning will be much slower but should lead to similar STDP features. Note that it seems that monkeys can recognize high level objects at scales and positions that have not been experienced previously (Hung et al., 2005; Logothetis et al., 1995). It could be that in the brain local learning and adaptive complex pooling are used up to a certain level of complexity, but not for high level objects. These high level objects could be represented with a combination of simpler features that would be already shift and scale invariant. As a result there would be less need for spatially specific representations for high level objects.

The third simplification we have made is to use only five layers (including the classification layer) whereas processing in the ventral stream involves many more layers (probably about ten), and complexity increases more slowly than suggested here. It could be that the additional depth of the primate visual system is entirely related to the need to have receptive field size increase. As mentioned above our system achieves position and size invariance

by weight sharing over large neural maps – a technique that cannot be used in the brain, where neurons probably pool from local slightly shifted receptive fields (see Chapter 5). However STDP as a way to combine simple features into more complex representations, based on statistical regularities among earliest spike patterns, seems to be a very efficient learning rule and could be involved at all stages.

The last main simplification we have made is to ignore both feed-back loops and top down influences. While normal, everyday vision extensively uses feed-back loops, the temporal constraints almost certainly rules them out in an ultra-rapid categorization task (Thorpe et al., 1996). The same cannot be said about the top-down signals, which do not depend directly on inputs. For *e.g.* there is experimental evidence that the selectivity to the 'relevant' features for a given recognition task can be enhanced in IT (Sigala and Logothetis, 2002) and in V4 (Bichot et al., 2005), possibly thanks to a top-down signal coming from the prefrontal cortex, thought to be involved in the categorization process. These effects, for example modeled in Szabo et al's model (Szabo et al., 2006), are not taken into account here.

2.6.4 'Early vs. later spike' coding and STDP: two keys to understand fast visual processing

Despite these four simplifications we think our model captures two key mechanisms used by the visual system for rapid object recognition. The first one is the importance of the first spikes for rapidly encoding the most important information about a visual stimulus. Given the number of stages involved in high level recognition and the short latencies (∼100 ms) of selective responses recorded in monkeys' IT (Oram and Perrett, 1992; Keysers et al., 2001; Hung et al., 2005), the time window available for each neuron to perform its computation is probably around 10-20 ms (Thorpe and Imbert, 1989) and will rarely contain more than one or two spikes (see Section 1.2.2). The only thing that matters for a neuron is whether or not an afferent fires early enough so that the presynaptic spike falls in the critical time window, while later spikes can not be used for ultra-rapid categorization. At this point (but only at this point) we have to consider two hypotheses: either presynaptic spike times are completely stochastic (for example, drawn from a Poisson distribution), or they are somewhat reliable. The first hypothesis causes problems since the first presynaptic spikes (again the only ones taken into account) will correspond to a subset of the afferents that is essentially random, and will not contain much information about their real excitation (Gautrais and Thorpe, 1998). A solution to this problem is to use populations of redundant neurons

(with similar selectivity), to ensure the first presynaptic spikes do correspond on average to the most active populations of afferents. In this work we took the second hypothesis, assuming the time-to-first spike of the afferents (or to be precise their firing order) was reliable and did reflect a level of excitation. This second hypothesis receives experimental support (see Section 1.5 for a review).

Very interestingly STDP provides an efficient way to develop selectivity to first spike patterns, as shown in this Chapter. After convergence the potential reached by a STDP neuron is linked to the number of early spikes in common between the current input and a stored prototype. This 'early spike' vs. 'later spike' neural code (while the spike order within each bin does not matter) has not only been proven robust enough to perform object recognition in natural images but is fast to readout: an accurate response can be produced when only the earliest afferent have fired. The use of such a mechanism at each stage of the ventral stream could account for the phenomenal processing speed achieved by the visual system.

Recently (Rolls et al., 2006) recorded neurons in IT and expressed some skepticism about the use of some sort temporal coding to encode object identity in this area. However our model is actually fully compatible with what they found. First they conclude that "considerable information is available from the first spike to arrive in response to a stimulus" – we obviously agree on that; second that "the order in which the spikes arrive from the different neurons does not appear to add significant information to that available from knowing that a spike has arrived from some but not other neurons" – which corresponds exactly the 'early spike' vs. 'later spike' we propose; and third that "more information is available if all the spikes in a short time window are taken into account" – we certainly agree that more information is available if we integrate spikes over a larger time window (25-50 ms in their case) (otherwise why would there be more than one spike?), but we argue that because of the above mentioned temporal constraints, in an ultra-rapid visual categorization task most of the processing if probably done with only the first spikes.

2.7 Technical details

Here is a detailed description of the network, the STDP model and the classification methods.

2.7.1 S_1 cells

S_1 cells detect edges by performing a convolution on the input images. We are using 5x5 convolution kernels, that roughly correspond to Gabor filters with wavelength of 5 (*i.e.* the kernel contains one period), effective width 2, and four preferred orientations: $\pi/8$, $\pi/4 + \pi/8$, $\pi/2 + \pi/8$ and $3\pi/4 + \pi/8$ ($\pi/8$ is there to avoid focusing on horizontal and vertical edges, that are seldom diagnostic). We apply those filters to five scaled versions of the original image: 100%, 71%, 50%, 35% and 25%[1]. There are thus 4 x 5 = 20 S_1 maps. S_1 cells emit spikes with a latency that is inversely proportional to the absolute value of the convolution (the response is thus invariant to an image negative operation). We also limit activity at this stage: at a given processing scale and location only the spike corresponding to the best matching orientation is propagated.

2.7.2 C_1 cells

C_1 cells propagate the first spike emitted by S_1 cells in a 7x7 square of a given S_1 map (that correspond to one preferred orientation and one processing scale). Two adjacent C_1 cells in a C_1 map correspond to two 7x7 squares of S_1 cells shifted by 6 S_1 cells (and thus overlap of one S_1 row). C_1 maps thus sub-sample S_1 maps. To be precise, neglecting the side effects, there are 6 x 6 = 36 times fewer C_1 cells than S_1 cells. As proposed by Riesenhuber and Poggio (Riesenhuber and Poggio, 1999), this maximum operation is a biologically plausible way to gain local shift invariance. From an image processing point of view, it is a way to perform sub-sampling within retinotopic maps without flattening high spatial frequency peaks (as would be the case with local averaging).

We also use a local lateral inhibition mechanism at this stage: when a C_1 cell emits a spike, it increases the latency of its neighbors within a 11x11 square in the map with the same preferred orientation and the same scale. The percentage of latency increase decreases linearly with the distance from the spike: 15% at 1 pixel, 12.5% at 2, 10% at 3, 7.5% at 4, and to 5% at 5 pixels. As a result, if a region is clearly dominated by one orientation, cells with inhibit each other and the spike train will be globally late so unlikely to be 'selected' by STDP.

[1] Choosing the number of scales is always an issue: the more scales the more likely you are that the actual scale of a target object do correspond to one processing scale, but more scales also generates more false alarm. A geometric progression with ratio $\sqrt{2}$ seemed a good choice.

2.7.3 S_2 cells

S_2 cells correspond to intermediate complexity visual features. Here we used ten prototype S_2 cell types, and twenty in the mixed simulation. Each prototype cell is duplicated in 5 maps (weight sharing), each map corresponding to one processing scale. Within those maps the S_2 cells can only integrate spikes from the four C_1 maps of the corresponding processing scale. The receptive field size is 16 x 16 C_1 cells (neglecting the side effects, this leads to 96 x 96 S_1 cells and the corresponding receptive field size in the original image is (96/processing scale)[2]). C_1-S_2 synaptic connections are set by STDP.

Note that we did not use a leakage term. In the brain, by progressively resetting membrane potentials towards their resting levels, leakiness will decrease the interference between two successive spike waves. In our model we process spike waves one by one, and reset all the potentials before each propagation, and so leaks are not needed. This is equivalent to studying the limit case when the membrane time constant is $\gg$ than the time-lag between two input spikes of the same wave, but is $\ll$ than the time-lag between two waves.

Finally, activity is limited at this stage: a k-Winner-Take-All ensures that at most two cells can fire for each processing scale. This mechanism, only used in the learning phase, helps the cells to learn patterns with different real sizes. Without it, there is a natural bias towards 'small' patterns (*i.e.* large scales), simply because corresponding maps are larger, so the likelihood of firing with random weights at the beginning of the STDP process is higher.

2.7.4 C_2 cells

Those cells take for each prototype the maximum response (*i.e.* first spike) of corresponding S_2 cells over all positions and processing scales, leading to ten shift and scale invariant cells (twenty in the mixed case)[2].

2.7.5 STDP Model

We used a simplified STDP rule:

$$\Delta w_{ij} = \begin{cases} a^+ \cdot w_{ij} \cdot (1 - w_{ij}) & \text{if} \quad t_j - t_i \leq 0 \quad \text{(LTP)} \\ a^- \cdot w_{ij} \cdot (1 - w_{ij}) & \text{if} \quad t_j - t_i > 0 \quad \text{(LTD)} \end{cases} \tag{2.3}$$

[2]Ten C_2 cells is a minimum: tests with fewer cells did not reach the same level of performance, because the variability of the target class was not covered.

where i and j refers respectively to the post- and pres-synaptic neurons, t_i and t_j are the corresponding spike times, Δw_{ij} is the synaptic weight modification, and a^+ and a^- are two parameters specifying the amount of change. Note that the weight change does not depend on the exact $t_j - t_i$ value, but only on its sign. We also used an infinite time window. These simplifications are equivalent to assuming that the intensity-latency conversion of S_1 cells compresses the whole spike wave in a relatively short time interval (say 20-30 ms), so that all presynaptic spikes necessarily fall close to the postsynaptic spike time, and the change decrease becomes negligible. In the brain this change decrease and the limited time window are crucial: they prevent different spike waves coming from different stimuli from interfering in the learning process. In our model, we propagate stimuli one by one, so these mechanisms are not needed. Note that with this simplified STDP rule only the *order* of the spikes matters, not their precise timings. As a result the intensity-latency conversion function of S_1 cells has no impact, any monotonously decreasing function give the same results.

The multiplicative term $w_{ij} \cdot (1 - w_{ij})$ ensures the weight remains in the range $[0, 1]$ (excitatory synapses), and implements a soft bound effect: when the weight approaches a bound, weight changes tend towards zero.

We also applied LTD to synapses through which no presynaptic spike arrived, exactly as if a presynaptic spike had arrived after the postsynaptic one. This is useful to eliminate the noise due to original random weights on synapses through which presynaptic spike never arrive.

As the STDP learning progresses we increase a^+ and $|a^-|$. To be precise we start with $a^+ = 2^{-6}$, multiply the value by 2 every 400 postsynaptic spikes, until it reaches a maximum value of 2^{-2}. Similarly a^- is adjusted so as to keep a fixed a^+/a^- ratio (-4/3). This allows us to accelerate convergence when the preferred stimulus is somewhat 'locked', whereas directly using high learning rates with the random initial weights leads to erratic results.

We used a threshold of 64 (=1/4 of the 16 x 16 $C_1 - S_2$ weights). Initial weights are randomly generated, with mean 0.8 and standard deviation 0.05.

2.7.6 Classification setup

We used a Radial Basis Function (RBF) network. In the brain this classification step may be done in the Pre-Frontal Cortex (PFC) using the outputs of IT. Let X be the vector of C_2 responses (containing either binary detections with the first implementation or final graded potentials with the second one).

This kind of classifier computes an expression of the form:

$$f(X) = \sum_{i=1}^{N} c_i \cdot \exp\left(-\frac{(X - X_i)^2}{2\sigma^2}\right) \tag{2.4}$$

and then classifies based on whether or not $f(X)$ reaches a threshold. Supervised learning at this stage involves adjusting the synaptic weights c_i so as to minimize a (regularized) error on the training set (Poggio and Bizzi, 2004).

$$E(X) = \frac{1}{2}\sum_{i=1}^{N} [d_i - F(X_i)]^2 + \frac{1}{2}\lambda \|\mathbf{D}F\|^2 \tag{2.5}$$

where the X_i correspond to C_2 responses for some training examples (1/4 of the training set randomly selected), $d_i = 1$ for positive examples and -1 for negative ones, λ is the regularization parameter, and $\mathbf{D}$ is a linear differential operator (the second term thus penalizes non-smooth functions). The full training set was used to learn the c_i. We used $\sigma = 2$ and $\lambda = 10^{-12}$.

The multi-class case was handled with a 'one versus all approach'. If n is the number of classes (here three), n RBF classifiers of the kind 'class i' vs. 'all other classes' are trained. At the time of testing each one of the n classifier emits a (real valued) prediction that is linked to the probability of the image belonging to its category. The assigned category is the one that correspond to the highest prediction value.

2.7.7 Hebbian learning

The spike trains coming from C_1 cells were converted into real valued activities (supposed to correspond to firing rates) by taking the inverse of the first spikes' latencies (note that these activities do not correspond exactly to the convolution values because of the local lateral inhibition mechanism of layer C_1). The activities (or firing rates) of S_2 units were computed using Equation 2.1. Note that the normalization causes an S_2 cell to respond maximally when the input vector X_{C_1} is collinear to its weight vector W_{S_2} (neural circuits for such normalization have been proposed in (Poggio and Bizzi, 2004)). Hence W_{S_2} (or any vector collinear to it) is the preferred stimulus of the S_2 cell. With another stimulus the response is proportional to the cosine between W_{S_2} and X_{C_1}. This kind of tuning has been used in extensions of HMAX (Serre et al., 2005b, 2007). It is similar to the Gaussian tuning of the original HMAX (Riesenhuber and Poggio, 1999), but it is invariant to the norm of the input (*i.e.* multiplying the input activities by 2 has no effect on the response) which allow us to remain contrast-invariant (see also (Serre

et al., 2005a) for a comparison between the two kinds of tuning).

Only the cells whose activities were above a threshold were considered in the competition process. It was found useful to use individual adaptive thresholds: each time a cell was among the winners its threshold was set to 0.91 times its activity (this value was tuned to get approximately the same number of weight updates than with STDP)[3]. The competition mechanism was exactly the same as before, except that it selected the most active units, and not the first one to fire. The winners' weight vector were updated with the modified hebbian rule of Equation 2.2, in which a is the learning rate. It was found useful to start with a small learning rate (0.002) and geometrically increase it every 10 iterations. The geometric ratio was set to reach a learning rate of 0.02 after 2000 iterations after which the learning rate stayed constant.

2.7.8 Differences from the model of Serre, Wolf and Poggio

Here we summarize the differences between our model and the model of (Serre et al., 2005b) in terms of architecture (leaving the questions of learning and temporal code aside)

- We process various scaled versions of the input image (with the same filter size), instead of using various filter sizes on the original image. This is equivalent, while being much faster.
- S_1 level: only the best matching orientation is propagated. This was shown to give better results in their model as well (Mutch and Lowe, 2006).
- C_1 level: we use lateral inhibition (see above) to avoid focusing on zones where one orientation dominates.
- S_2 level: the similarity between a current input and the stored prototype is linked to the number of early spikes in common between the corresponding spike trains, while Serre *et al.* use the Euclidian distance between the corresponding patches of C_1 activities.
- We used an RBF network and not a Support Vector Machine (although this had little impact on the performance).

[3]This kind of threshold, which essentially says that a certain level of activity has to be reached for the synapses to be plastic, differs from the sliding threshold of the Bienenstock Cooper Munro (BCM) theory (Bienenstock et al., 1982), which determines whether LTP or LTD is applied. Note that STDP has recently been mapped to a BCM learning rule (Toyoizumi et al., 2005; Pfister and Gerstner, 2006)

Chapter 3

STDP-based spike pattern learning

The study presented in this Chapter has just been published (Masquelier et al., 2008):

Masquelier T, Guyonneau R, Thorpe SJ (2008) Spike Timing Dependent Plasticity Finds the Start of Repeating Patterns in Continuous Spike Trains. *PLoS ONE* 3(1): e1377. doi:10.1371/journal.pone.0001377

The original paper can be found on Section A.2.

3.1 Résumé

De nombreuses études expérimentales ont observé une Long Term Potentiation (LTP) quand un neurone pré-synaptique décharge peu de temps avant un neurone post-synaptique, et une Long Term Depression (LTD) quand le neurone pré-synaptique décharge peu de temps après, un phénomène connu sous le nom de Spike Timing Dependent Plasticity (STDP) (Bi and Poo, 1998; Markram et al., 1997; Zhang et al., 1998; Feldman, 2000; Vislay-Meltzer et al., 2006; Mu and Poo, 2006; Cassenaer and Laurent, 2007).

De nombreuses études théoriques ont modélisé les effets de la STDP quand les spikes en entrée arrivent par *vagues* successives. On montre alors que, sous réserve que ces vagues présentent des similarités, la STDP concentre les poids synaptiques sur les afférents qui sont systématiquement parmi les premiers à décharger, ce qui a pour effet diminuer la latence du spike post-synaptique par rapport au début de la vague (Song et al., 2000; Gerstner and Kistler, 2002; Guyonneau et al., 2005; Masquelier and Thorpe, 2007). La STDP a

également été étudiée en régime oscillatoire, et il a été montré qu'elle permet de sélectionner uniquement les entrées dont la phase est constante parmi une population d'afférents aux phases aléatoires (Gerstner et al., 1996).

Les effets de STDP en régime continu sont plus méconnus. Ici, on s'intéresse au cas très général d'un neurone qui est face à une population d'afférents qui déchargent continuellement avec un taux moyen uniforme et constant. On montre que, étonnamment, même dans cette situation STDP est capable de détecter puis de remonter un pattern de spikes spatio-temporel qui se répète, pourtant 'caché' dans des trains de spikes 'distracteurs' de même densité, ce qui est un problème computationnellement complexe (Frostig et al., 1990; Prut et al., 1998; Abeles and Gat, 2001; Fellous et al., 2004). A la fin de l'apprentissage, le neurone est devenu sélectif au début du pattern et peut servir de prédicteur précoce pour le reste du pattern, au risque de produire une fausse alarme si le reste du pattern ne vient pas, mais avec l'avantage d'être très réactif.

STDP permet donc l'utilisation d'un codage temporel, même sans date de référence explicite. Cela signifie que des discontinuités globales comme les saccades en vision, les sniffs en olfaction (Uchida et al., 2006), ou les oscillations en général ne sont pas nécessaires pour l'apprentissage basé sur STDP, bien qu'ils le facilitent probablement. Etant donné que le mécanisme présenté ici est simple et peu coûteux, il est difficile de croire que le cerveau n'a pas évolué pour l'utiliser.

3.2 Abstract

Experimental studies have observed Long Term synaptic Potentiation (LTP) when a presynaptic neuron fires shortly before a postsynaptic neuron, and Long Term Depression (LTD) when the presynaptic neuron fires shortly after, a phenomenon known as Spike Timing Dependent Plasticity (STDP). When a neuron is presented successively with discrete volleys of input spikes STDP has been shown to learn 'early spike patterns', that is to concentrate synaptic weights on afferents that consistently fire early, with the result that the postsynaptic spike latency decreases. Here, we show that these results still stand in a continuous regime where afferents fire continuously with a constant population rate. As such, STDP is able to solve a very difficult computational problem: to localize a repeating spatio-temporal spike pattern embedded in equally dense 'distractor' spike trains. STDP thus enables some form of temporal coding, even in the absence of an explicit time reference. Given that the mechanism exposed here is simple and cheap it is hard to believe that the brain did not evolve to use it.

3.3 Introduction

3.3.1 The computational problem: spike pattern detection

Electrophysiologists report the existence of repeating spatio-temporal spike patterns with millisecond precision, both *in vitro* and *in vivo*, lasting from a few tens of ms to several seconds (Frostig et al., 1990; Prut et al., 1998; Fellous et al., 2004; Ikegaya et al., 2004; Abeles, 2004). In this study we assess the difficult problem of detecting a repeating spatio-temporal pattern in multiple spike trains, a problem made particularly difficult when only a fraction of the recorded neurons are involved in the pattern. Fig. 3.1 illustrates such a situation. There is a pattern of spikes (indicated by the red dots) that repeats at irregular intervals, but is hidden within the variable background firing of the whole population (shown in blue). The problem is made hard because nothing in terms of population firing rate characterizes the periods when the pattern is present, nor is there anything unusual about the firing rates of the neurons involved in the pattern. In such a situation detecting the pattern clearly requires taking the spike times into account. However direct comparison of each spike time to one another over the entire recording period and across the entire set of afferents is extremely computationally expensive. However, in this article we will see how a single neuron equipped with STDP can solve the problem in a different manner, taking advantage of the fact that a pattern is a succession of spike coincidences.

3.3.2 Background: STDP and discrete spike volleys

STDP is now a widely accepted physiological mechanism of activity-driven synaptic regulation (see Section 1.7.1 for experimental evidence). Note that STDP is in agreement with Hebb's postulate because it reinforces the connections with the presynaptic neurons that fired slightly before the postsynaptic neuron, which are those which 'took part in firing it'. It thereby reinforces causality links.

When a neuron is presented successively with similar volleys of input spikes STDP is known to have the effect of concentrating synaptic weights on afferents that consistently fire early, with the result that the postsynaptic spike latency decreases (Song et al., 2000; Gerstner and Kistler, 2002; Guyonneau et al., 2005; Masquelier and Thorpe, 2007). This theoretical observation is in accordance with recordings in rat's hippocampus showing that the so called 'place cells' fire earlier – relative to the cycle of the theta oscillation in hippocampus – after the animal has repeatedly traversed the correspond-

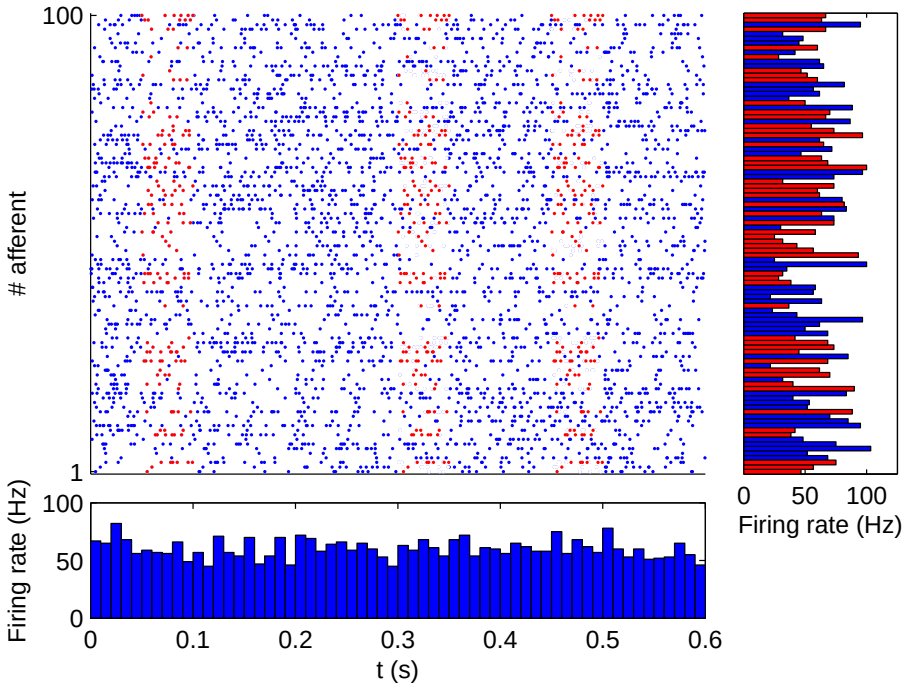


Figure 3.1: Spatio-temporal spike pattern. Here we show in red a repeating 50 ms long pattern that concerns 50 afferents among 100. The bottom panel plots the population spike counts (in spikes) using 10 ms time bins (we chose 10 ms because it is the membrane time constant of the neuron used later in the simulations), and demonstrates that nothing in terms of spike count characterizes the periods when the pattern is present. The right panel plots the individual spike counts over the whole period. Neurons involved in the pattern are shown in red. Again, nothing characterizes them in terms of spike count. Detecting the pattern thus requires taking the spike times into account.

ing area (Mehta et al., 2000). STDP has also been studied in an oscillatory mode, and was shown to be able to select only phase-locked inputs among a broad population with random phases, turning the postsynaptic neuron into a coincidence detector (Gerstner et al., 1996).

The main limitation of these studies is the assumption that the input spikes arrive in discrete volleys (sometimes also called 'spike waves'). They assume an explicit time reference – usually the presentation of a stimulus (Song et al., 2000; Guyonneau et al., 2005; Masquelier and Thorpe, 2007), or the maximum (or minimum) of an oscillatory drive (Hopfield, 1995; Gerstner et al., 1996) – that allows the specification of a time-to-first spike (or latency) for the afferents, which could be used by the brain to encode information (VanRullen et al., 2005; Fries et al., 2007). Activity between the volleys is assumed to be spontaneous and much weaker. Furthermore, many studies (Gerstner et al., 1996; Song et al., 2000; Guyonneau et al., 2005) also require the pattern to be present in all volleys for the STDP to learn it, that is no 'distractor' volleys are inserted between pattern presentations. But what happens when the population of afferents is continuously firing with a constant population firing rate, so that no explicit time reference is available? Is STDP still able to find and learn spike patterns among the inputs? Is the learning robust if, more realistically, pattern presentations occur at unpredictable times, separated by long 'distractor' periods and if the pattern does not involve all the afferents? Does it make sense to use the beginning of the pattern as a time reference, and does the postsynaptic spike latency with respect to this reference still decrease?

3.3.3 Experimental set-up: STDP in continuous regime

To answer these questions we inserted an arbitrary pattern at various times into randomly generated 'distractor' spike trains, as in Fig. 3.1, and investigated whether a single receiving STDP neuron, with a 10 ms membrane time constant, was able to learn it in an unsupervised manner. To be precise, we simulated a population of 2,000 afferents firing continuously for 450 s (see Section 3.6 for details). Most of the time (3/4 of the time in the baseline simulation) the afferents fired according to a Poisson process with variable instantaneous firing rates. Spiking activity in the brain is usually assumed to follow roughly Poisson statistics, hence this choice, but here it is not crucial: what matters is that the afferents fire stochastically and independently. But every now and then, at random times, half of these afferents left the stochastic mode for 50 ms and adopted a precise firing pattern. This repeated pattern had roughly the same spike density as the stochastic distractor part, so as to make it invisible in terms of firing rates. To be precise the firing

rate averaged over the population and estimated over 10 ms time bins has a mean of 64 Hz and a standard deviation of less than 2 Hz (this firing rate is even more constant than in the 100 afferent case of Fig. 3.1 because of the law of large numbers). We further increased the difficulty by adding a permanent 10 Hz Poissonian spontaneous activity to all the neurons, and by adding a 1 ms jitter to the pattern. Intriguingly, we will see that one single Leaky Integrate-and-Fire (LIF) neuron receiving inputs from all the afferents, acting as a coincidence detector (see Fig. 1.5), and implementing STDP, is perfectly able to solve the problem and learns to respond selectively to the start of the repeating pattern.

3.4 Results

3.4.1 A first example

At the beginning of a first simulation the 2,000 synaptic weights all equal to 0.475 (arbitrary units normalized in the range [0,1]). The neuron is therefore non-selective. Since the presynaptic spike density – on its 10 ms time scale – is almost constant, it discharges periodically (see Fig. 3.2a). The greater are the initial weights (or the lower the threshold), the smaller is the period (here it is about 16 ms). Each time a discharge occurs we update the synaptic weights using the STDP rule of Fig. 1.6, and clip them in the range [0,1]. At this stage, the neuron discharges both outside and inside the pattern (represented by grey rectangles on Fig. 3.2). In the first case presynaptic and postsynaptic spike times are uncorrelated, and since $a^-\tau^- > a^+\tau^+$ (LTD/LTP imbalance; see Section 3.6), STDP leads to an overall weakening of synapses (Song et al., 2000) (note: if no repeating patterns were inserted STDP would thus gradually decrease the synaptic weights until the threshold would not be reached any longer). But in the second case, by reinforcing the synaptic connections with the afferents that took part in firing the neuron, STDP increases the probability that the neuron fires again next time the pattern is presented (reinforcement of causality link). As a result, selectivity to the pattern emerges, here after about 13.5 s (see Fig. 3.2b) that is after only about 70 pattern presentations and 700 discharges: the neuron gradually stops discharging outside the pattern (no false alarms), while it does discharge most of the time when the pattern is presented (high hit rate), and can even fire twice per pattern as in the case illustrated here. Chance determines which part(s) of the pattern the neuron becomes selective to at this stage (*i.e.* the postsynaptic spike latency(ies), with respect to the beginning of the pattern here about 5 ms and 40 ms). However the increase

in selectivity usually rapidly leads to only one discharge per pattern, here at about 40 ms.

Once selectivity to the pattern has emerged STDP has another major effect. Each time the neuron discharges in the pattern, it reinforces the connections with the presynaptic neurons that fired slightly before in the pattern. As a result next time the pattern is presented the neuron is not only more likely to discharge to it, but it will also tend to discharge earlier. In other words, the postsynaptic spike latency locks itself to the pattern and decreases steadily (with respect to the beginning of the pattern). However, it cannot decrease endlessly. There is a convergence by saturation when all the spikes in the pattern that precede the postsynaptic spike already correspond to maximally potentiated synapses, and all are necessary to reach the threshold. This usually occurs when the latency is already very short, the value depending on the threshold, although it could occur even earlier if the pattern has a zone with low spike density. Spikes outside the pattern cannot contribute efficiently to the membrane potential: since their times are stochastic, STDP usually depresses the corresponding synapses. We end up with a bimodal weight distribution with synapses either maximally potentiated or fully depressed (as predicted by VanRossum et al. (2000)).

Here this convergence occurs after about 2000 discharges. At this stage, the postsynaptic spike latency (with respect to the beginning of the pattern) is about 4 ms (see Fig. 3.2c). After convergence the hit rate is then 99.1% with no false alarms (estimated on the last 150 s). Notice that the signal/noise ratio has increased with respect to the situation in Fig. 3.2b, that is the potential reached on distractor periods is farther from the threshold. Among the 2,000 synapses, 383 are fully potentiated (weight$\approx$1), while the rest of them are almost completely depressed (weight$\approx$0). All of the potentiated synapses correspond to afferents involved in the pattern. The fact that there is no false alarms means once the learning has been done, a neuron just waits for its preferred stimulus, and need never forget what it has learned. The model thus predicts that fully specified neurons might actually have very low spontaneous rates, whereas higher rates might characterise less well specified cells.

Fig. 3.3 shows the latency reduction (with respect to the beginning of the pattern) during the learning stage until it stabilizes at a minimum of about 4 ms. Apart from the initial part (before selectivity emerges) the curve looks similar to those observed in earlier work with discrete spike volleys (Guyonneau et al., 2005). By convention the latency is 0 when the neuron discharged outside the pattern, that is when it generated a false alarm. There are no false alarms after the 676th discharge, that is for the last 436 s of simulation.

Fig. 3.4 illustrates the situation after convergence. It can be seen that

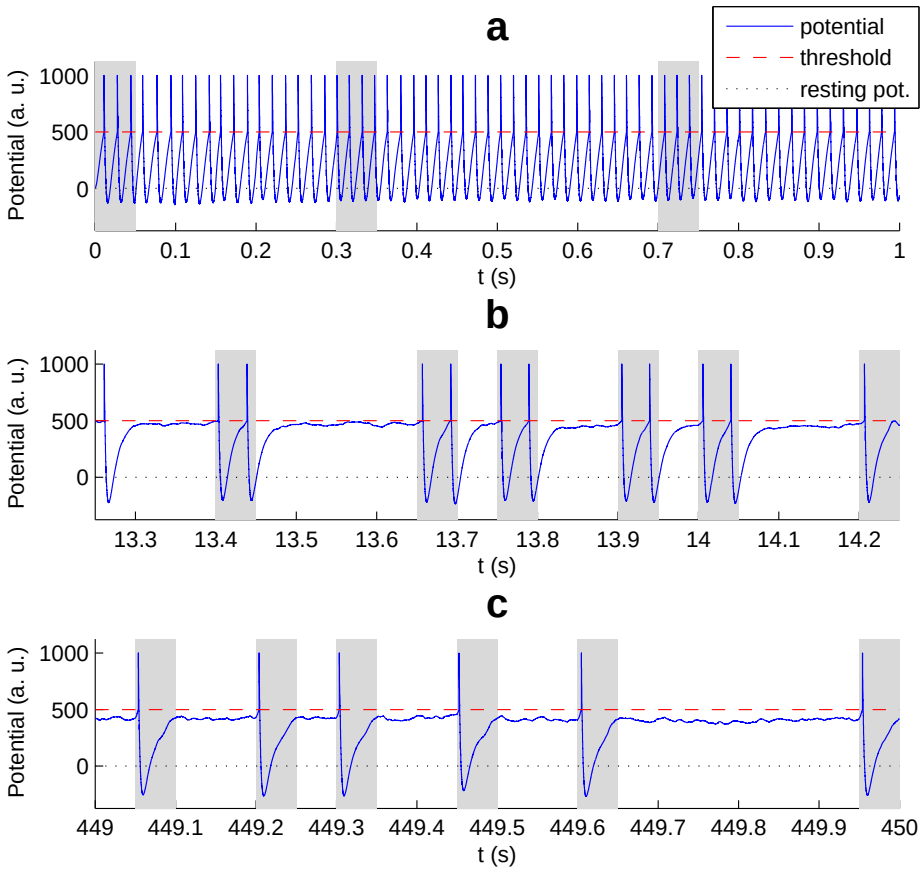


Figure 3.2: Overview of the 450 s simulation. Here we plotted the membrane potential as a function of simulation time, at the beginning, middle, and end of the simulation. Grey rectangles indicate pattern presentations. (a) At the beginning of the simulation the neuron is non-selective because the synaptic weights are all equal. It thus fires periodically, both inside and outside the pattern. (b) At t=13.5 s, after about 70 pattern presentations and 700 discharges, selectivity to the pattern is emerging: gradually the neuron almost stops discharging outside the pattern (no false alarms), while it does discharge most of the time the pattern is present (high hit rate), here even twice (c) End of the simulation. The system has converged (by saturation). Postsynaptic spike latency is about 4 ms. Hit rate is 99.1% with no false alarms (estimated on the last 150 s).

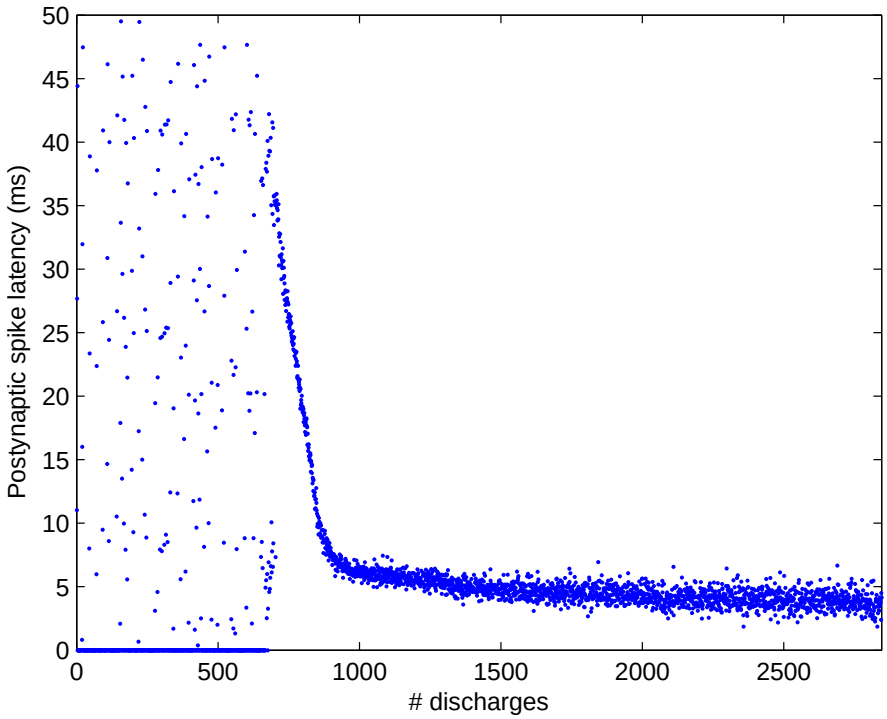


Figure 3.3: Latency reduction. Here we plotted the postsynaptic latency as a function of the number of discharges (by convention the latency is 0 when the neuron discharged outside the pattern, *i.e.* when it generated a false alarm). We clearly distinguish 3 periods: the beginning, when the neuron is non-selective; the middle, when selectivity has emerged and STDP is 'tracking back' through the pattern; and the end, when the system has converged towards a fast and reliable pattern detector.

STDP has potentiated most of the synapses that correspond to the earliest spikes of the pattern (Fig. 3.4a), and depressed most of the synapses that correspond to presynaptic spikes which follow the postsynaptic one, as in the previous work with discrete volleys (Song et al., 2000; Guyonneau et al., 2005; Masquelier and Thorpe, 2007). This results in a sudden increase in membrane potential when the neuron starts integrating the pattern, and the threshold is quickly reached (Fig. 3.4b). Notice that all the synaptic connections with afferents not involved in the pattern have been completely depressed.

3.4.2 Batches

We performed 100 similar simulations with different pseudo-randomly generated spike trains. Our criteria for a 'successful' simulation were: convergence to a state with a postsynaptic latency inferior to 10 ms, a hit rate superior to 98% and no false alarms. This occurred in 96% of the cases. For the remaining 4%, the neurons stopped firing when too many discharges occurred outside the pattern in a row (leading to an overall weakening of synapses, so the threshold was no longer reached).

We ran other batches of 100 simulations to investigate the impact on this 96% success performance of five parameters.

The first one is the pattern relative frequency (i.e sum of pattern durations over total duration ratio, assuming a fixed pattern duration of 50 ms), 1/4 in the baseline condition, and Fig. 3.5a shows its effect. We see that while the performance is very high as long as the ratio is above 15%, with smaller values the probability of success slowly drops. This means the pattern needs to be consistently present for the STDP to learn it. However, this applies only at the beginning (say during the first 1000 discharges). Here we used a constant pattern frequency, but after the initial part the neuron has already become selective to the pattern, so presenting longer distractor periods does not perturb the learning at all. We also tried to change the pattern duration while maintaining its relative frequency at 1/4. It turns out that what makes the detection difficult is the delay between two pattern presentations, not the pattern duration itself. Since we kept the pattern relative frequency constant, this delay increased with the pattern duration so the performance dropped: 97% with a 40 ms pattern, 96% with 50 ms, 93% with 60 ms, 59% with 100 ms and 46% with 150 ms. However we think this delay is more naturally investigated by changing the pattern relative frequency as in Fig. 3.5a.

The second parameter we investigated is the amount of jitter (1 ms in the baseline condition), and Fig. 3.5b shows its influence. We see that the performance is very good for jitter levels lower than 3 ms. For larger amounts

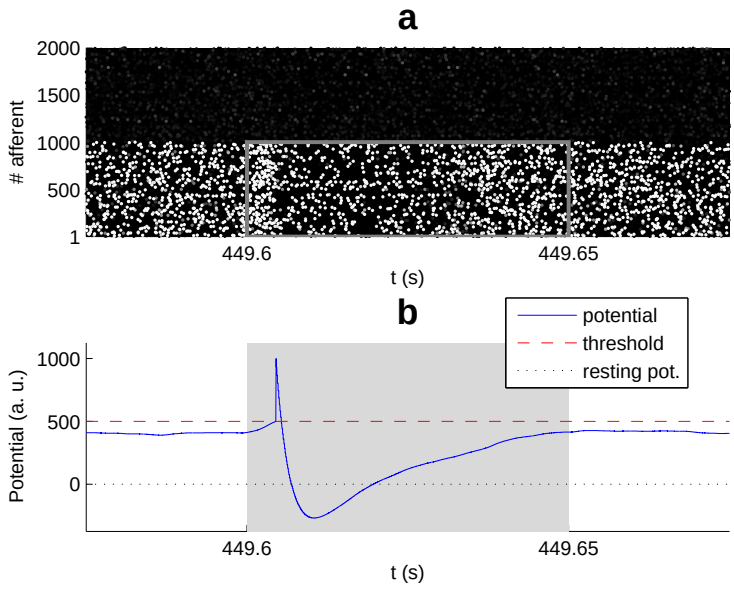


Figure 3.4: Converged state (a) we represented the spike trains of the 2,000 afferents. We have reordered the afferents with respect to Fig. 3.1 so that afferents 1-1000 are involved in the pattern, and afferents 1001-2000 are not and we use a color code ranging from black for spikes that correspond to completely depressed synapses (weight=0) to white for spikes that correspond to maximally potentiated synapses (weight=1). This allows the visualization of the spikes which generate a significant EPSP and those which do not. The pattern is represented with a grey line rectangle. Notice the cluster of white spikes at the beginning of it: STDP has potentiated most of the synapses that correspond to the earliest spikes of the pattern. Note that virtually all the synaptic connections with afferents not involved in the pattern have been completely depressed. (b) The membrane potential is plotted as a function of time, over the same range as above. We clearly see the sudden increase that corresponds to the above-mentioned cluster.

of jitter the spike coincidences (with respect to the 10 ms membrane time constant) are lost, and the STDP weight updates are inaccurate, so the learning is impaired. In the brain millisecond spiking precision has been reported in many structures (see Section 1.5).

The third parameter is the proportion of afferents involved in the pattern (1/2 in the baseline condition), and Fig. 3.5c shows its influence. The threshold was scaled proportionally. Not surprisingly, with fewer afferents involved in the pattern, it becomes harder to detect, but it is still detected more than half of the times when only 1/3 of the afferents are involved in the pattern. Note that the other 2/3 of afferents are discarded by STDP. This suggests that activity-driven mechanisms could select a small set of 'interesting' afferents among a much bigger set of initially connected afferents, probably specified genetically, a phenomenon known as 'developmental exuberance' for which there is considerable experimental evidence (Innocenti and Price, 2005).

The fourth parameter is the initial weight (0.475 in the baseline condition) and Fig. 3.5d shows its influence. Recall discharges outside the pattern lead to an overall decrease of synaptic weights. If too many of them occur in a row the threshold may no longer be reachable. Thus a high initial value for the weights increases the resistance to discharges outside the pattern, leading to a better performance. High initial weights also cause the neuron to discharge at a high rate at the beginning of the learning process, when it is non-selective: 63 Hz for an initial weight of 0.475, 38 Hz for 0.325. These values may seem high in regard to usual experimental values. But first after only 13 s selectivity has emerged, and the neuron fires at a rate between 5 and 10 Hz. It is conceivable that electrophysiologists rarely record such short very active initial phases. Second, we consider here that the population of afferents is constantly firing with a mean rate of 64Hz. This is to make the problem of pattern detection harder, but if the afferents have less active periods, which is likely to occur in the brain, so will have the post-synaptic neuron. We also added Gaussian noise to the initial weights, with increasing standard deviation until 0.475 (thus equal to the mean). Following this noise addition the weights were clipped in [0,1]. This had no significant impact on the performance, at least in the present case when the initial weights are relatively high.

The fifth parameter is the proportion of missing spikes (0 in the base line condition) and Fig. 3.5e shows its influence. Not surprisingly the number of successfully learned patterns decreases with the proportion of spikes deleted. However with a 10% deletion the pattern was correctly learnt 82% of the time, demonstrating that the system is quite robust to spike deletion.

We also tried changing the membrane time constant τ_m (10 ms in the

baseline condition), scaling the threshold proportionally. This had little impact on the performance (79% success with τ_m=5 ms, 88% with τ_m=20 ms), but it did have an impact on the minimal latency that is reached after convergence. A smaller time constant (and the smaller threshold that goes with it) causes the neuron to be interested in more coincident spikes. The system converges when the very few nearly coincident first spikes of the pattern all correspond to maximally potentiated synapses, and the postsynaptic spikes is fired just after them. The final latency is thus shorter than the one we have with a longer time constant, which enables the neuron to integrate spikes over a longer time window.

Taken together these results demonstrate that the learning is amazingly robust to the model parameters. We thus believe that we have captured a mechanism than emerges from STDP rather than from a precise neural model configuration. While we admit it is still somewhat speculative to affirm that a similar mechanism takes place in the brain, it is at least very plausible.

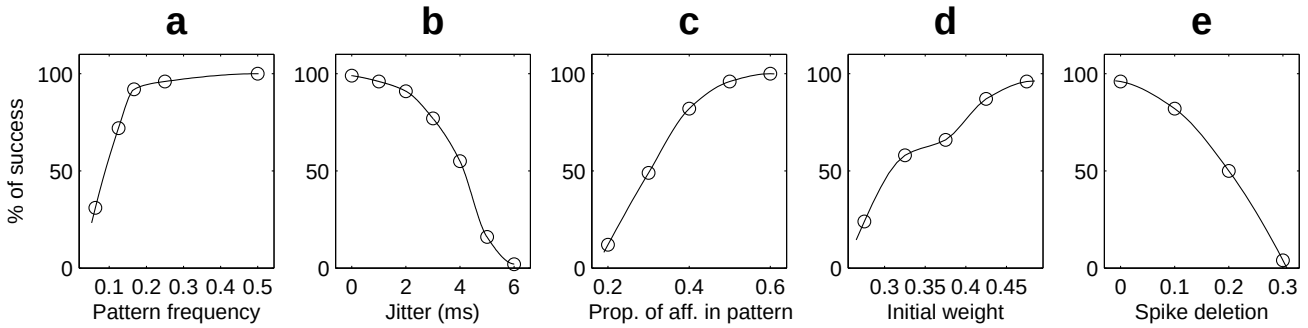


Figure 3.5: Resistance to degradations (100 trials). (a) Percentage of successful trials as a function of the pattern frequency (pattern duration / the total duration, given a fixed pattern length of 50 ms). The pattern needs to be consistently present, at least at the beginning, for the STDP to start the learning process. (b) Percentage of successful trials as a function of jitter. For jitter greater than 3 ms (this should be compared to the 10 ms membrane time constant) spike coincidences are lost and learning is impaired (c) Percentage of successful trials as a function of the proportion of afferents involved in the pattern. Performance is good if this proportion is above $1/3$ (d) Percentage of successful trials as a function of the initial weights. With a high value the neuron can handle more discharges outside the pattern. (e) Percentage of successful trials as a function of the proportion of spikes deleted. With a 10% deletion the pattern was correctly learnt in 82% of the cases.

3.5 Discussion

3.5.1 STDP in continuous regime

Our first claim is that the main results previously obtained for STDP based learning with the highly simplified scheme of discrete spike volleys (Song et al., 2000; Gerstner and Kistler, 2002; Guyonneau et al., 2005; Masquelier and Thorpe, 2007) still stand in this more challenging continuous framework. This means that global discontinuities such as saccades or micro-saccades in vision and sniffs in olfaction (Uchida et al., 2006), or brain oscillations in general are not necessary for STDP-based learning of temporal patterns, although they will almost certainly help. Temporal code skeptics often point out the fact that neurons would need to know a time reference to decode a temporal code, and we see here that this is not necessary: as long as there are recurrent spike patterns in the inputs, and even if they are embedded in equally dense 'distractor' spike trains, a neuron equipped with STDP can potentially find them in only a few tens of pattern presentations, and will gradually respond faster and faster when the pattern is presented, by potentiating synapses that correspond to the earliest spikes of the patterns, and depressing all the others. This last point strongly reinforces the idea that a substantial amount of information could be available very rapidly, in the very first spikes evoked by a stimulus (VanRullen and Thorpe, 2001). Our results also suggest that activity-driven mechanisms could select a small set of afferents among a much bigger set of initially connected afferents, probably specified genetically, a phenomenon known as 'developmental exuberance' for which there is considerable experimental evidence (Innocenti and Price, 2005).

It is worth mentioning that the proposed learning scheme is fully unsupervised. No teaching signal tells the neuron when to learn nor labels the inputs. Biologically plausible mechanisms for supervised learning of spike patterns have also been proposed (Gütig and Sompolinsky, 2006).

3.5.2 Spike pattern detection

It is also surprising to see how such a simple mechanism can solve a problem as complex as spike pattern detection. However, there is no consensus on the definition of a spike pattern, and we admit ours is quite simple: here a pattern is seen as a succession of coincidences. A Leaky Integrate and Fire (LIF) neuron is known to be capable of coincidence detection, and it has even been proposed that this is its main function in the brain (Abeles, 1982; Konig et al., 1996). Here the membrane time constant (10 ms) is shorter than the

duration of the pattern (50 ms), and so the LIF neuron can never be selective to the whole pattern. Instead, it is selective to 'one coincidence' of the pattern at a time, that is, selective to the nearly simultaneous arrival of certain spikes, just as it occurs in one subdivision of the pattern. At the beginning of the learning process STDP will cause the LIF neuron to become selective to one such coincidence (chance determines which one). Then STDP will track back through the pattern, from one coincidence to the previous one, until the initial coincidence is reached and the chain of causality is stopped. At this point the neuron is selective only to the simultaneous arrival of the pattern's earliest spikes, and can serve as 'earliest predictor' of the subsequent spike events (Song et al., 2000; Mehta et al., 2000; Gerstner and Kistler, 2002), at the risk of triggering a false alarm if these subsequent events don't occur, but with the benefit of being very reactive.

This contrasts with approaches where the whole pattern needs to be taken into account, sometimes including finer structural aspects such as spike orders or relative delays (Frostig et al., 1990; Prut et al., 1998; Abeles and Gat, 2001; Fellous et al., 2004). But neuronal mechanisms able to reliably decode such structures have to be proposed and looked for in the brain. One appealing candidate mechanism is the synfire chain (Abeles, 1991) but direct evidence for their existence is still fairly limited (Abeles, 2004). Here we limit the notion of pattern to successive coincidences, and suggest a way such patterns could be decoded, using widely accepted neurophysiological mechanisms, namely coincidence detection and STDP.

Another limitation of this work is the excitatory-only scheme. Consequently, something like 'afferent A must not spike' cannot be learnt, only 'positive patterns' can. However, evidence for plasticity in inhibitory synapses in the brain is weak and inhibition is often assumed to be non-selective. So we propose that most of the selectivity could be achieved using only excitatory synapses, as in this model.

3.5.3 Argument for temporal coding

Whether spike times contain additional information with respect to discharge rates has been the object of an ongoing debate for some time. Electrophysiologists have tried to answer this question mostly by recording neurons in sensory and motor systems with a repeating stimulus or action, and looking at inter-trial variability of the spike times. Some claim that spike times can be very reliable while others are more skeptical (see Section 1.5 for a review). Given that the simple and cheap mechanism exposed here reliably detects spatio-temporal spike patterns, it is hard to believe that the brain did not evolve to use at least the form of temporal coding exposed above ('succes-

sive coincidences'), unless there is an unavoidable intrinsic source of noise in the integrate-and-fire mechanism that makes all spike times unreliable. The main source for this sort of noise is probably at the level of synaptic transmission (Movshon, 2000), since neurons stimulated directly by current injection in the absence of synaptic input give highly stereotyped and precise responses (Mainen and Sejnowski, 1995). However, spike times can be very reliable in some experiments (see Section 1.5), particularly in the auditory cortex, proving that reliable synapses do exist. So, as said in Section 1.4.2, we argue that variability in other recorded spike times, in particular in the visual system, could come from non-controlled variables that might also affect neuronal activation, such as attention, eye movements, mental imagery, top-down effects *etc.*

3.5.4 A generic mechanism

We would like to emphasize the fact that the approach presented here is generic. It is not limited to sensory systems, and it could be applied to either experimental or model-generated data. The first step would be to see if STDP finds spike patterns in the data. Providing it does, the second step would be to understand what those patterns mean by solving the corresponding inverse problem.

3.5.5 Extension: competitive scheme

What happens if there is more than one repeating pattern present in the input? We verified that as the learning progresses, the increasing selectivity of the postsynaptic neuron rapidly prevents it from responding to several patterns. Instead, it picks one (chance determines which one), and becomes selective to it and only to it. To learn the other patterns other neurons are needed.

A competitive mechanism could ensure they optimally cover all the different patterns and avoid learning the same ones. Such a mechanism could be implemented through inhibitory horizontal connections between neurons, such that as soon as one neuron fires, it could prevent other cells from learning the same pattern, as in previous work (Guyonneau et al., 2004). The neural population would then self-organize to cover all the input patterns. The 'coverage' could be optimized using neurons that differ in their parameters (for example their thresholds), leading to more robust learning and detection. Furthermore a long input pattern can be coded by the successive firings of several STDP neurons, each selective to a different part of the pattern, and competition would prevent them all from tracking back through the

pattern and clustering at the beginning. Note that within such a competitive framework a pattern detection probability of 50% is hardly a disaster: it means that with 2 neurons the risk that one pattern is not detected is 25%, with 3 neurons 12.5%, with 4 neurons 6.25% and so on. The system could then work with suboptimal parameters (highlighted in Fig. 3.5), for example weaker initial weights.

Further work is needed to evaluate this form of competitive network. However in this Chapter we wanted to stress the fact that one single LIF neuron equipped with STDP is consistently able to detect one arbitrary repeating spatio-temporal spike pattern embedded in equally dense 'distractor' spike trains, which is a remarkable demonstration of the potential for such a scheme.

3.6 Technical details

The simulations were performed using MATLAB R14 (Mathworks 2005, Natick MA). The source code is available upon request.

3.6.1 Poisson spike trains

The spike trains were prepared before the simulation (Fig. 3.1 illustrates the type of spike trains we used, though with a smaller set of neurons). For memory issues instead of using spike trains defined over a 450 seconds period, we pasted the same 150 s long pattern three times (this repetition had no impact on the results). Each afferent emits spikes independently using a Poisson process with a variable instantaneous firing rate r, that varies randomly between 0 and 90 Hz. The maximal rate change s was chosen so that the neuron could go from 0 to 90 Hz in 50 ms. To be precise, time was discretized using a time step dt of 1 ms. At each time step:

1. the afferent has a probability of r.dt of emitting a spike (whose exact date is then picked randomly in the 1 ms time bin)

2. its instantaneous firing rate is modified: dr = s.dt where s is the speed of rate change (in Hz/s), and clipped in [0, 90] Hz.

3. its speed of rate change is modified by ds, randomly picked from a uniform distribution over [-360 +360] Hz/s, and clipped in [-1800 +1800] Hz/s

Note that we chose to apply the random change to s as opposed to r so as to have a continuous s function and a smoother r function.

As mentioned in the Discussion, a limitation of this work is the excitatory-only scheme. Consequently, something like 'afferent A must not spike' cannot be learnt, only 'positive patterns' can. We thus wanted a pattern in which all the afferents spike at least once. We could have made up such a pattern, but we wanted the pattern to have exactly the same statistics as the Poisson distractor part (to make the pattern detection harder), so we preferred to randomly pick a 50 ms period of the original Poisson spike trains and to 'copy-paste' it (see below). To make sure this randomly selected period did contain a spike from each afferent we implemented a mechanism that triggers a spike whenever an afferent has been silent for more than 50 ms (leading to a minimal firing rate of 20 Hz). Clearly, such mechanism is NOT implemented in the brain. It is just an artifice we used here to make the pattern detection harder. As a result the average firing rate was 54 Hz, and not the 45 Hz we would have without this additional mechanism.

Once the random spike train has been generated, a part of it, defined as the 'pattern' to be repeated, is 'copy-pasted'. This 'copy-paste' does not involve the last 1000 afferents (obviously the indices are arbitrary), which conserve their original spike trains. But we discretize the spike trains of the first 1000 afferent into 50 ms sections. We randomly pick one of these sections and copy the corresponding spikes. Then we randomly pick a certain number of these sections (1/4 in the baseline condition), avoiding consecutive ones, and replace the original spikes by the copied ones. A jitter was added before the pasting operation, picked from a Gaussian distribution with mean zero and standard deviation 1 ms (in the baseline condition).

After this 'copy-paste' operation a 10 Hz Poissonian spontaneous activity was added, to all neurons and all the time. The total activity was thus 64 Hz on average, and spontaneous activity represented about 16% of it.

3.6.2 Leaky Integrate and Fire (LIF) neuron

(see Fig. 1.5) For computational reasons we modeled the LIF neuron using Gerstner's Spike Response Model (SRM) (Gerstner, 1995; Gerstner and Kistler, 2002). That is instead of solving the membrane potential differential equation we used kernels to model the effect of presynaptic and postsynaptic spikes on the membrane potential. Each presynaptic spike j, with arrival time t_j, is supposed to add to the membrane potential an Excitatory Post-Synaptic Potential (EPSP) of the form:

$$\epsilon(t-t_j) = K\left(\exp\left(-\frac{t-t_j}{\tau_m}\right) - \exp\left(-\frac{t-t_j}{\tau_s}\right)\right)\cdot\Theta(t-t_j) \qquad (3.1)$$

where τ_m is the membrane time constant (here 10 ms), τ_s is the synapse time constant (here 2.5 ms), Θ is the Heavyside step function:

$$\Theta(s) = \begin{cases} 1 & \text{if} \quad s \geq 0 \\ 0 & \text{if} \quad s < 0 \end{cases} \tag{3.2}$$

and K is just a multiplicative constant chosen so that the maximum value of the kernel is 1 (the voltage scale is arbitrary in this Chapter).

The last emitted postsynaptic spike i has an effect on the membrane potential modeled as follows:

$$\begin{aligned} \eta(t - t_i) = \Theta(t - t_i) \cdot T \cdot \\ \left\{ K_1 \cdot \exp\left(-\frac{t-t_j}{\tau_m}\right) - K_2 \cdot \left(\exp\left(-\frac{t-t_i}{\tau_m}\right) - \exp\left(-\frac{t-t_i}{\tau_s}\right)\right)\right\} \end{aligned} \tag{3.3}$$

where T is the threshold of the neuron (here 500, arbitrary units). The first term models the positive pulse and the second one the negative spike-afterpotential that follows the pulse (see Fig. 1.5). Here we used $K_1 = 2$ and $K_2 = 4$. For simplicity, the resting potential is supposed to be zero, but a non zero value would simply shift the kernel, and shifting the threshold by the same value would lead to the same computation.

Both ϵ and η kernels were rounded to zero when respectively $t - t_j$ and $t - t_j$ were greater than $7\tau_m$.

At any time the membrane potential is:

$$p = \eta(t - t_i) + \sum_{j|t_j > t_i} w_j \epsilon(t - t_j) \tag{3.4}$$

where the w_j are the excitatory synaptic weights, between 0 and 1 (arbitrary units).

This SRM formulation allows us to use event-driven programming: we only compute the potential when a new presynaptic spike is integrated. We then estimate numerically if the corresponding EPSP will cause the threshold to be reached in the future and at what date. If it is the case, a postsynaptic spike is scheduled. Such postsynaptic spike events cause all the EPSPs to be flushed, and a new t_i is used for the η kernel. There is then a refractory period of 1 ms, during which the neuron is not allowed to fire.

This event driven programming is much less computationally expensive than solving numerically the LIF differential equation using a small time step.

3.6.3 Spike Timing Dependent Plasticity

An exponential update rule (see Fig. 1.6):

$$\Delta w_j = \begin{cases} a^+ \cdot \exp\left(\frac{t_j - t_i}{\tau^+}\right) & \text{if} \quad t_j - t_i \leq 0 \quad \text{(LTP)} \\ a^- \cdot \exp\left(-\frac{t_j - t_i}{\tau^-}\right) & \text{if} \quad t_j - t_i > 0 \quad \text{(LTD)} \end{cases} \tag{3.5}$$

with the time constants $\tau^+ = 16.8$ ms and $\tau^- = 33.7$ ms, provides a reasonable approximation of the synaptic modification observed experimentally (Bi and Poo, 2001).We restricted the learning window to $[t_i - 7\tau^+, t_i]$ for LTP and to $[t_j, t_j + 7\tau^-]$ for LTD. For each afferent, we also limited LTP (respectively LTD) to the last (first) presynaptic spike before (after) the postsynaptic one ('nearest spike' approximation). We did not take the effects of finer triplet of spikes (Pfister and Gerstner, 2006) into account.

It was found that small learning rates led to more robust learning. We used $a^+ = 2^{-5}$ and $a^- = -0.85a^+$. Following learning the weights were clipped to [0,1]. Note that all synapses remain excitatory: there is no inhibition in all these simulations.

Chapter 4

Visual learning experiment

In this Chapter I present recent unpublished psychophysical data. The experiment was run at the Centre de Recherche Cerveau et Cognition, Toulouse, France.

4.1 Résumé

Une prédiction des modèles STDP présentés aux Chapitres 2 et 3 est que la latence des réponses neuronales devrait décroître après répétition d'un même stimulus. Ceci devrait être vrai en particulier dans le cortex visuel qui semble rester plastique durant toute la vie (voir Section 1.3). C'est là une prédiction intéressante, et qui peut être testée expérimentalement.

Si l'électrophysiologie fournirait un moyen direct de tester cette prédiction, ce n'est pas l'approche que nous avons utilisée ici : nous avons choisi d'inférer les temps de traitements visuels en utilisant une mesure comportementale. Plus précisément nous avons utilisé le paradigme de choix forcé saccadique (Kirchner and Thorpe, 2006; Guyonneau et al., 2006; Bacon-Macé et al., 2007; Fletcher-Watson et al., 2007; Girard et al., 2007), dans lequel une image cible et un distracteur sont flashés simultanément de part et d'autre d'une croix de fixation, et le sujet doit diriger son regard vers l'image cible le plus rapidement possible. On enregistre à la fois la précision (taux de réponses correctes) et les temps de réaction. Ici on répéte toujours la même image cible (une scène d'intérieur), versus des distracteurs changeant (également des scènes d'intérieur), et on cherche à savoir si les temps de réactions diminuent au long de l'expérience.

Effectivement, on a constaté que les temps de traitement s'accéléraient d'environ 100 ms, même si ce gain est en majeure partie imputable à l'apprentissage de la tâche, et seulement 25 ms est spécifique à l'image. Ces 25 ms

sont gagnées au bout de quelques centaines de présentations. Beaucoup de questions restent ouvertes, notamment si cet effet de 25 ms dépend du type de stimuli utilisé, et s'il est invariant à la position rétinienne, c'est à dire : reconnaîtrions nous aussi vite une image apprise à une position donnée si elle nous est présentée à un autre endroit ?

Bien sûr l'existence d'un effet d'accélération par apprentissage ne prouve pas que les modèles STDP présentés aux Chapitres 2 et 3 sont vrais – ils sont seulement plausibles.

4.2 Abstract

One of the predictions of STDP models of Chapters 2 and 3 is that visual responses' latencies should decrease after repeated presentations of a same stimulus. In this Chapter we tested this prediction experimentally by inferring the visual processing times through behavioral measures.

We used the saccadic forced-choice paradigm (Kirchner and Thorpe, 2006; Guyonneau et al., 2006; Bacon-Macé et al., 2007; Fletcher-Watson et al., 2007; Girard et al., 2007), in which one target image and one distractor image are flashed simultaneously on both sides of a fixation cross, and the participant is asked to move his eyes towards the target as fast as possible. Both accuracy (*i.e.* correct response rate) and reaction times are recorded. Here the target was always the same repeating image (an interior scene), while the distractors (other interior scenes) were changing, and we looked for a familiarity-induced speed-up effect.

The experiment did reveal a familiarity-induced speed-up effect of about 100 ms. Most of it can be attributed to the learning of the task but a 25 ms effect corresponds to the familiarity with a given image, and is reached after a few hundred presentations. Many questions remain open, such as if this 25 ms speed-up is shift-invariant and depends on the type of stimuli.

4.3 Introduction

One of the predictions of STDP models of Chapters 2 and 3 is that neuronal responses' latencies should decrease after repeated presentations of a same stimulus. This should be true in particular in the visual cortex, that seems to be plastic even in adulthood (see Section 1.3 for evidence). It is an interesting prediction that is worth being tested experimentally.

Although electrophysiology would be a way to test directly the prediction, it is not the approach we took here: we chose to infer the neural processing

times using behavioral measures. To be precise we used the saccadic forced-choice paradigm (Kirchner and Thorpe, 2006; Guyonneau et al., 2006; Bacon-Macé et al., 2007; Fletcher-Watson et al., 2007; Girard et al., 2007), in which one target image and one distractor image are flashed simultaneously on both sides of a fixation cross, and the participant is asked to move his eyes towards the target as fast as possible. Both accuracy (*i.e.* correct response rate) and reaction times are recorded. This paradigm has been shown to reveal subtle differences in processing times between for example categories – differences that were invisible with a go/no-go task with manual response (for example face/non-face categorization is faster than animal/non-animal with the saccadic forced-choice paradigm (unpublished data) whereas the speed is about the same with the go/no-go paradigm (Rousselet et al., 2003b), Simon Thorpe, personal communication). Here the target was always the same repeating image, while the distractors were changing, and we looked for a familiarity-induced speed-up effect.

In a previous study Fabre-Thorpe et al. (2001) looked for an eventual experience-induced speed-up effect, using the animal/non-animal go/no-go paradigm of Thorpe et al. (1996). An extensive training with 200 animal images over a 3-week period failed to increase the speed of processing. This could be due to the fact that, as mentioned above, go/no-go is not always appropriate to reveal subtle differences. However Kirchner and Thorpe (2006) also looked for a image-specific learning effect with the saccadic forced-choice by repeating 20 target images for 20 times each. They found that there was a tendency for the reaction times to decrease during the course of the experiment (first vs. last block of 80 trials: 27 ms), but this general learning effect did not interact with the type of image present (new vs. repeated) and was therefore probably due to improvement in decision or motor skills. It thus seems we are already experts in animal/non-animal classification.

We thus needed a visual task in which we were not already expert. Ideally we wanted the participants to learn a visual stimulus that they had never seen before. A preliminary experiment where non-Chinese speaker participants were asked to learn and detect one Chinese character among others in a saccadic forced-choice paradigm showed that the task was too difficult. We thus used natural images, to be precise house interior scenes, and the participant was asked to identify one of them *among other interior scenes*. It is likely that before the experiment participants already had neurons tuned to features suitable for the encoding of interior scenes and the objects present in them (presumably located in the Para-hippocampal Place Area (PPA) and the Lateral Occipital Complex (LOC)). However the identification of a precise previously unseen scene among others requires specific learning. Whether this learning is done by building a 'grand-mother cell' that inte-

grates spikes from the appropriate feature detectors, or by other forms of read-out is unknown. In any case, the processes involved in this ultra-rapid visual categorization task are likely to be forward (Kirchner and Thorpe, 2006) and STDP should be able to speed them up.

Although many questions remain open the experiment did reveal such a speed-up effect. Most of it seems to be independent of the target image, that is participants seem to learn how to do the task, and develop strategies that are in fact useful for the detection of any interior scenes (for example focus on the gist or on a precise zone). However some of it ($\sim$25 ms) is image dependent, *i.e.* results from the familiarity with one given image, but is lost if a new target image is used.

4.4 Methods

4.4.1 Participants

Twelve volunteers (aged from 23 to 52, 5 women and 7 men) with normal or corrected-to-normal vision performed a 2AFC visual discrimination task. The experimental procedures were authorized by the local ethical committee (CCPPRB No. 9614003). Experiments were undertaken with the understanding and written consent of each participant.

4.4.2 The saccadic forced-choice

We used a modified version of the saccadic forced-choice paradigm originally proposed by Kirchner and Thorpe (2006). Fig. 4.1 illustrates the protocol. The background was a 50% gray. The participant had first to fix a cross shown in the middle of the screen for a pseudo-random fixation period P chosen in [800 ms, 1600 ms]. An eye tracker[1] monitored the gaze position in real time. When the participant had been fixing the cross continuously for P ms (with precision $\pm 1.7^\circ$), the cross was removed and two images were presented for 80 ms (unlike in (Kirchner and Thorpe, 2006), we did not use a gap between the cross removal and the image presentation). Each image corresponded to $8.4 \times 8.4^\circ$ of visual angle, and the eccentricity was $\pm 8.7^\circ$. The images were followed by two fixation crosses indicating the saccade landing positions. A time-out of 1 s was used for the participant's response. An inter-trial of 1.5 s followed. The participant was asked to move

[1]Model: iView X at 240 Hz, by SensoMotoric Instruments GmbH (SMI), Teltow/Berlin, Germany. I wrote the code to interface it with the Matlab PsychoToolBox presentation software.

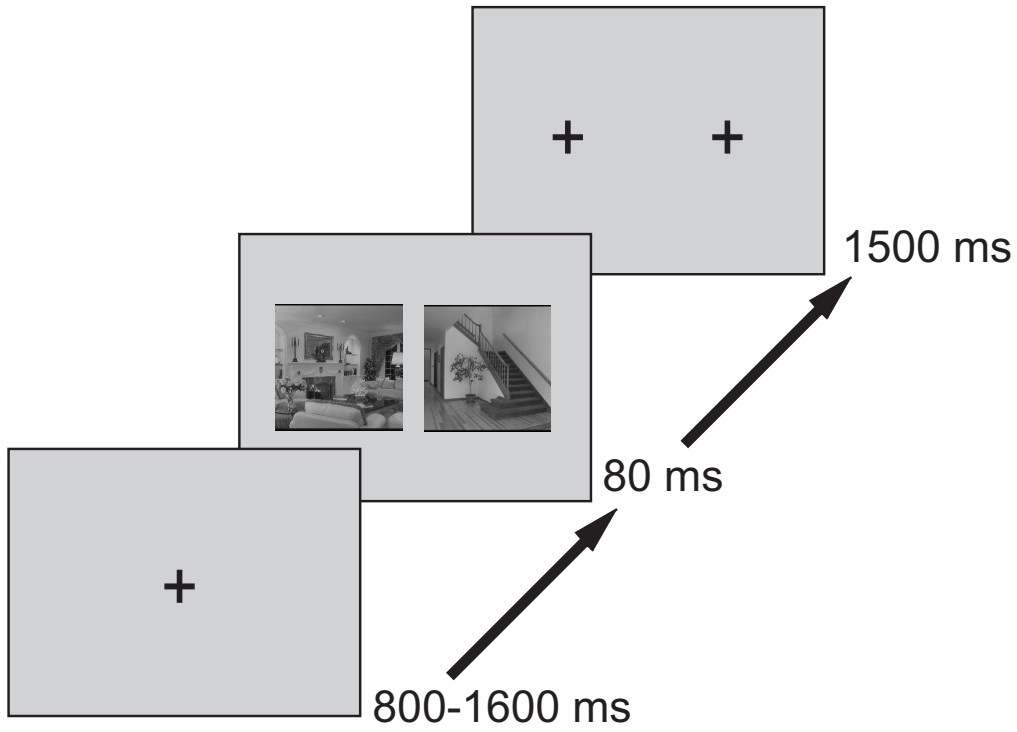


Figure 4.1: Saccadic forced-choice protocol. We checked that the participant did fixate the cross for a pseudo random period in [800 ms,1600 ms]. We then removed the cross and presented the two stimuli for 80 ms. The images were followed by two fixation crosses indicating the saccade landing positions.

his eyes towards the cross that corresponded to the target image. Note that the 80 ms presentation almost never allowed the participant to explore the images, that were already removed when he initiated his saccades in all but .2% of the cases.

4.4.3 Design

Each participant did two sessions, separated by 24 hours.

On Day 1 Participant 1 learnt two targets A and B. We used 12 alternated blocks of 50 trials. That is in each block the same target (A for even blocks, B for odd blocks) was shown 50 times in a row, versus changing distractors. The first two blocks began with two trials without distractor, that is only the target was shown on one side for 80 ms (nothing on the opposite side) so that the participant could identify it. These 'easy trials' are not taken into

account in the data analysis.

On Day 2 participant was tested on A and a new image C. The protocol was exactly the same as on the first day.

We chose a two-day protocol because we had reasons to believe that the learning of the repeated target would be consolidated during sleep (Karni et al., 1994; Stickgold et al., 2000; Atienza et al., 2004; Walker et al., 2005; Censor et al., 2006). This effect was not found significative here (see Section 4.5).

The three target image A, B and C were counter-balanced across participants (see Table 4.1). Note that full counter balancing would actually require 24 subjects, to include cases where the repeated image does not correspond to the same block parity on Day 1 and Day 2. We assumed such counter balancing was not needed.

Table 4.1: Experimental design: counter-balancing images across participants.

Participant	D 1 - Even	D 1 - Odd	D 2 - Even	D 2 - Odd
1	A	B	A	C
2	B	A	B	C
3	C	A	C	B
4	A	C	A	B
5	B	C	B	A
6	C	B	C	A
7	B	A	C	A
8	A	B	C	B
9	A	C	B	C
10	C	A	B	A
11	C	B	A	B
12	B	C	A	C

4.4.4 Stimuli

A hundred and one 246×246 gray level images were used, all representing house interior scenes[2] (see Fig. 4.2 for sample images). Mean luminance was

[2]Source: Corel CD-ROM library, "Residential interiors"

Figure 4.2: Sample pictures from the database.

normalized to 43%. Contrast, defined as the standard deviation of the gray levels, was normalized to 16%.

Three of them were chosen to be the targets A, B and C (see Fig. 4.3). We chose images with no salient object on the foreground.

4.4.5 Saccade detection

Two computers were used: the presentation computer, and the eye-tracker computer. The first one dealt with the stimuli presentation (using the Matlab PsychoToolBox). The second one recorded the gaze position constantly. Each block began with a 13-point calibration procedure. At the time of flash-

Figure 4.3: The three target pictures A, B and C.

ing the images the presentation computer sent a trigger to the eye-tracker computer, through the ethernet port and a crossover cable, as recommended by SMI. This trigger appeared in the gaze position files and allowed to compute off-line the saccade initiation times with respect to the stimulus onset.

Saccades were detected as follows: first we determined the first time after stimulus onset that the gaze shifted by more than half the distance between the original fixation cross and the center of the images (shorter saccades were not considered as valid responses). Later points were not considered (so if the subject changed his mind after a large saccade we took the first choice into account). The initiation of the saccade was then defined as the last time the speed passed the threshold of $85° /s$. This allowed to define a saccade initiation time in more than 86% of the cases.

4.5 Results

To our surprise the task was found significantly more difficult than the original animal/non-animal task of Kirchner and Thorpe (2006) (see Section 4.6.3). This is the reason why we had to remove the 200 ms gap they used between the fixation cross and the image presentation to reach a similar accuracy, but at the expense of slower reaction times (see Section 4.6.4)[3]. Note that 2 out of 14 original participants were considered as outliers and were discarded: they did not reach 70% on this last period, but they were also the fastest with median reaction times of 171 and 142 ms. This suggests a bound on the speed-accuracy trade-off curve: under ∼200 ms, it is virtually impossible to do the task.

[3]In a pilot experiment with a 200 ms gap I was 58% accurate with a mean reaction time of 175 ms, vs. 71% correct and 250 ms without the gap.

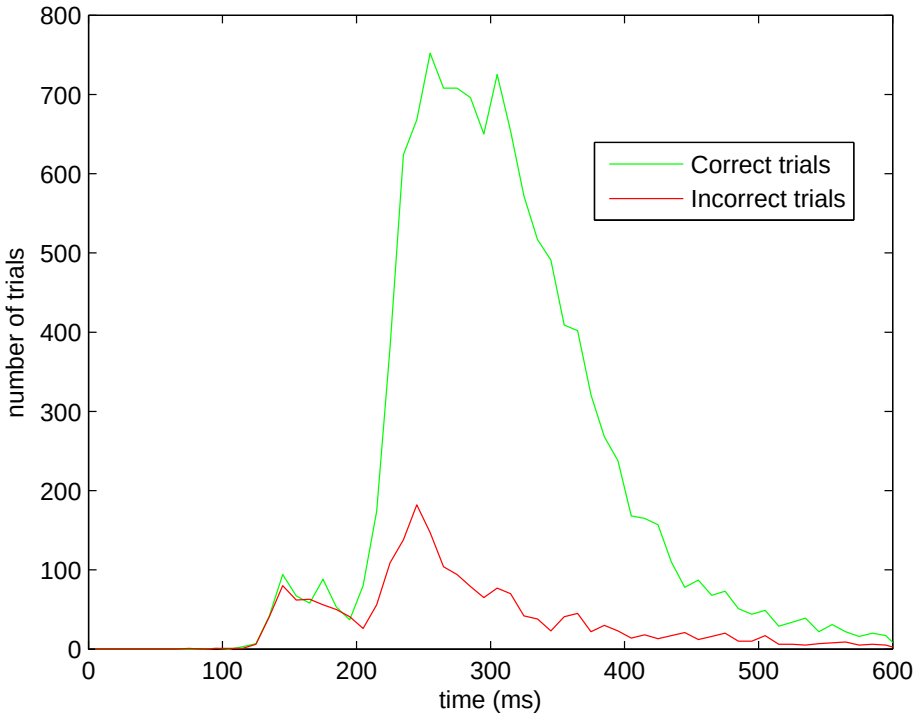


Figure 4.4: Histograms of reaction times, for both correct and incorrect trials, for all the participants and all the blocks. Note that responses under ∼200 ms are at chance level.

Fig. 4.4 plots the raw histograms of reaction times, for both correct and incorrect trials, for all the participants and all the blocks, and confirms this ∼200 ms bound under which participants are at chance level. On each session each target image was seen 300 times, that we split in 4 phases of 75 trials. We give the following name to our conditions: 1R for the target of Day 1 that will be repeated on Day 2, 1S for the single target of Day 1, 2R for the repeated target of Day 2 and 2S for the single target of Day 2.

Fig. 4.5 plots the speed-accuracy curves for one typical participant. Those curves are obtained from the raw distributions of the kind shown in Fig. 4.4 by computing the cumulative distribution of the correct minus incorrect responses, normalized by the number of trials (here 75). $y = 1$ thus corresponds to 100% correct and $y = 0$ to chance level (50% correct). The x corresponds to the time bound until which responses are taken into account. It can be seen that, despite some variability, the curves tend to climb faster and higher

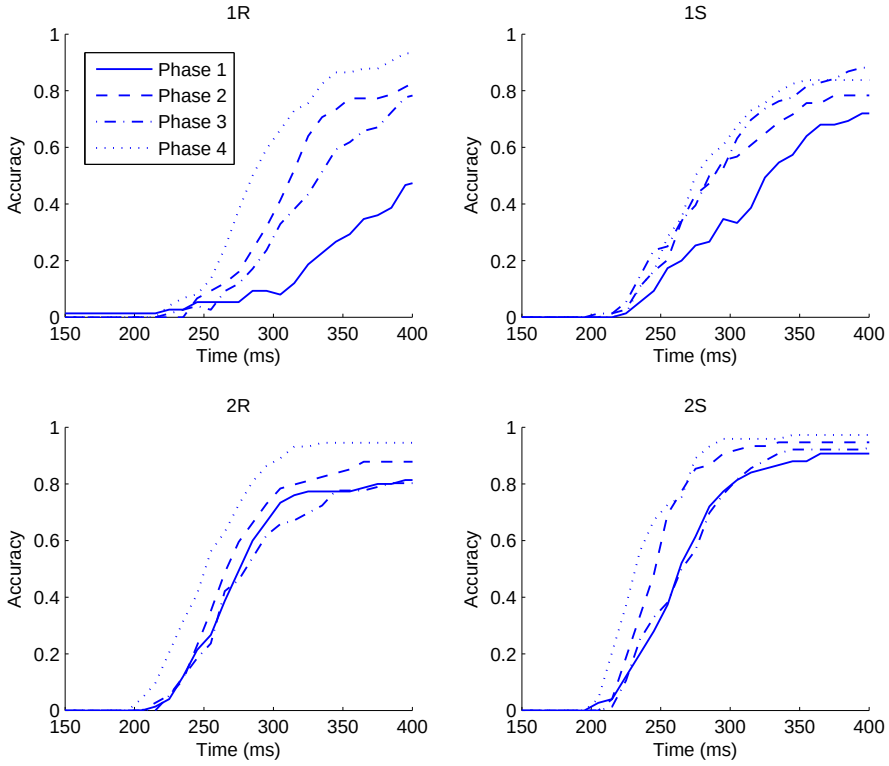


Figure 4.5: Speed-accuracy curves for one typical participant, for each of the four phases of 75 trials, in the four conditions 1R, 1S, 2R and 2S. It can be seen that, despite some variability, the curves tend to climb faster and higher for later phases, which is the signature of a learning effect.

for later phases, which is the signature of a learning effect. We summarize this phenomenon by computing the abscissas of the intersections of the curves with $y = 0.12$, that we call the 'reference times'. This evaluation criterion combines both the accuracy and the reaction times. The 12% threshold may seem low, but we are above all interested in the fast responses. When participants took more time to answer, they probably used higher cognitive process such as complex decision-making, presumably involving feedback, which led to high variability. In this study, we focus on the fast and more stereotyped responses, whose neuronal correlate is presumably mostly feedforward and more likely to be speeded-up by STDP.

We thus have 12 participants x 4 conditions x 4 phases = 192 data points. Four of them could not be computed, because the good answers never out-

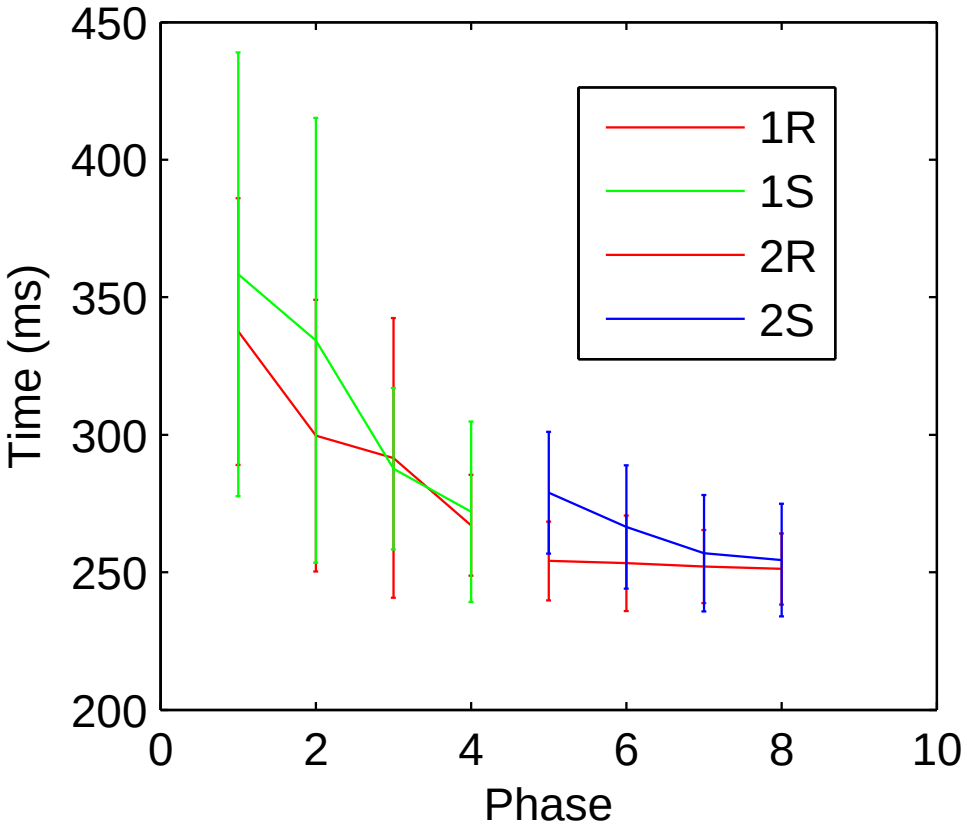


Figure 4.6: Reference times for the four phases and the four conditions, averaged over all participants.

numbered the errors by 12%. We thus replaced the each missing value by the maximum reference time of the corresponding participant, taken over all the phases of the corresponding condition. Note that this is a conservative bias that tend to minimize the learning effect.

Fig. 4.6 plots the reference times for the four phases and the four conditions, averaged over all participants. Note that the error bars are big, indicating a high inter-subject variability. This is the reason why we used paired statistical tests. Note also that reference times are really high and highly variable on Day 1, especially on the first two phases, when participants have to learn a difficult task. The apparent difference between 1R and 1S on those first two phases is presumably just a fluke. Day 2 data is much cleaner.

The first conclusion is that there is indeed a learning effect. References

times decrease from about 350 ms to 250 ms on average between the beginning and the end of the experiment. Differences between images were not found significant (1-way repeated measure ANalysis Of VAriance (ANOVA)). The second conclusion is that only about 25 ms of this speed-up is image dependent. Most of the speed-up can thus be attributed to the fact that the participants learn how to do the task, tune their motor systems, and develop strategies that are in fact transferable to the detection of another interior scenes (for example focus on the gist or on a precise zone).

We performed repeated measure ANOVA to check the significance of those effects. Given that most reaction times follow a log-normal distribution we first took the logarithm of the measured reference time. Before applying each ANOVA we also check for the sphericity of the data using both Bartlett's and Mauchly's sphericity tests, which both concluded that assumption of sphericity was tenable in all our case tests.

We performed a first repeated measure ANOVA with all the data, and three factors: Day, Repeated vs. Single (RvsS), Phase. Only the day, the phase, and the interaction of both were found significant explicative factors. This validates the learning effect but it does not seem to be image-specific. We expected the 'Repeated vs. Single' criterion to be a significant explicative factor when interacting with the day or the phase or both, but it was not the case. We attribute that to the high unexplained variability of the first phases of Day 1 which is not negligible with respect to the difference between repeated and single conditions on Day 2.

We thus performed a second repeated measure ANOVA with a subset of the data consisting of the last phase of Day 1 and the first phase of Day 2, and two factors: Day and Repeated vs. Single (RvsS) (see Table 4.2). This time, the interaction Day x RvsS has a significant effect on the reference times ($p=0.026$). This means the difference between the Repeated and the Single conditions on Day 2 is signifantly greater than the unexplained difference between them on Day 1. It thus validates the advantage for the image we know already.

We performed a third ANOVA with the Day 2 data, and two factors: Phase and Repeated vs. Single (RvsS)(see Table 4.3). The interaction Phase x RvsS has a significant effect on the reference times ($p=0.024$). This means that, as expected, the Repeated condition leads to significantly faster reference times, but only for the first phases. After only ~ 150 trials the new image is recognized as fast as the known one.

Table 4.2: ANOVA Day 1 last phase - Day 2 first phase. Factor 1: Day. Factor 2: Repeated vs. Single (RvsS). Legend: SOV: Source Of Variability, SS: Sum of Squares, df: degree of freedom, MS: Mean Square, F: F statitic, P: p-value.

SOV	SS	df	MS	F	P
Day	0.000	1	0.000	0.036	0.8528
Error(Day)	0.136	11	0.012		
RvsS	0.026	1	0.026	2.232	0.1633
Error(RvsS)	0.128	11	0.012		
DayxRvsS	0.021	1	0.021	6.646	0.0257
Error(DayxRvsS)	0.035	11	0.003		
Error	0.298	33	0.009		
Total	1.026	47			

Table 4.3: ANOVA Day 2. Factor 1: Phase. Factor 2: Repeated vs. Single (RvsS).

SOV	SS	df	MS	F	P
Phase	0.036	1	0.036	1.662	0.2238
Error(Phase)	0.240	11	0.022		
RvsS	0.038	3	0.013	2.503	0.0763
Error(RvsS)	0.167	33	0.005		
PhasexRvsS	0.025	3	0.008	3.587	0.0239
Error(PhasexRvsS)	0.077	33	0.002		
Error	0.483	77	0.006		
Total	1.510	95			

4.6 Discussion

4.6.1 A robust experience-induced speed-up

The difficulty we are facing with this experiment is that we have to estimate performance frequently to capture the learning effect, since it is fast ($\sim$ 100 presentations for the new image on Day 2). We thus used 75 trial bins, but such small bins lead to noisy estimations of the reference times.

However our results did show a significant 100 ms speed-up effect due to familiarity with both the task ($\sim$ 75 ms) and the target image ($\sim$ 25 ms). A few hundred presentations are needed to reach a ceiling effect. Note that we used a somewhat aggregated level, by summarizing 75 trials with one value, the so called 'reference time'. We thus believe our results are robust and reproducible.

Of course the existence of the speed-up does not prove the STDP models of Chapters 2 and 3 are true. It just shows that they are *plausible*.

4.6.2 Type of stimuli and shift-invariance

We are planning other experiments to investigate if this 25 ms effect is still there with other kinds of stimuli, such as random dots, fractals or Chinese characters.

Another interesting question is whether the 25 ms advantage for familiar images that we recorded is 'shift-invariant'. That is would we recognize faster a familiar object at a previously unseen retinal location? There are reasons to doubt it. Many perceptual learning studies, with task ranging from vernier discrimination to texture segmentation, report limited shift-invariance Ramachandran (1976); Fiorentini and Berardi (1981); Karni and Sagi (1991); Shiu and Pashler (1992); Fahle et al. (1995). Both Nazir and O'Regan (1990) and Dill and Fahle (1998) also showed that random patterns were harder to recognize at a previously unseen retinal location.

The answer probably depends on the type of stimuli. It is likely that high level stimuli, such as natural scenes, are encoded higher in the ventral stream hierarchy, possibly in IT, where responses are relatively invariant to translation. Simpler stimuli, such as random dots, might be encoded in lower areas, possibly in V1 or V2, where responses' position tuning is sharper.

We are planning future experiments to assess these questions.

4.6.3 Target-distractor distance has more impact than intra-class variability

As mentioned above the task was more difficult than expected. Participants reported that the task was demanding, especially at the beginning. They found hard that they never had a chance to see the target in foveal vision, let alone to explore it. It usually took them a few tenth of trials to identify the repeating target (the two 'easy trials' at the beginning with no distractor were usually not enough for the participants to identify the target). Even at the end of the experiment they could hardly describe the content of the targets.

What is surprising is that from a computational point of view the identification of a given scene seems much simpler than the recognition of a class with huge intra-class variability such as animals. A three line program could solve the first problem, whereas computer vision scientists are still fighting to solve the second one with human performance.

We attribute this difficulty to the fact that we used distractors of the same category as the target, namely interior scenes. Both must activate neurons in the same region(s) of the visual cortex and with the same latency(ies), thus a precise read-out procedure is probably needed to distinguish the target from the distractors. If on the contrary the distractors are from another category, as in the animal/non-animal task, they will probably not activate the same region(s), or at least not at the same time (see for example (Kiani et al., 2005), discussed in Section 1.5), so simply estimating the activity(ies) of the region(s) selective to the target, eventually at a given time, should be enough to do the task.

This idea is agreement with the go/no-go results of Delorme et al. (2004), who found a 10 ms advantage (with similar accuracy) for the identification of a given animal picture among non-animal distractors with respect to the identification of a given non-animal picture, still among non animal distractors. Although the 'non-animal' class is extremely varied, and we predict the 10 ms difference would increase using a more narrow category for both target and distractors, for example interior scenes like here, the difficulty of the task seems to be directly linked to similarity between target and distractors. Our results suggests that this similarity has more impact on the performance than the variability of the target class (zero in the extreme case of identification). This is summarized in Fig. 4.7.

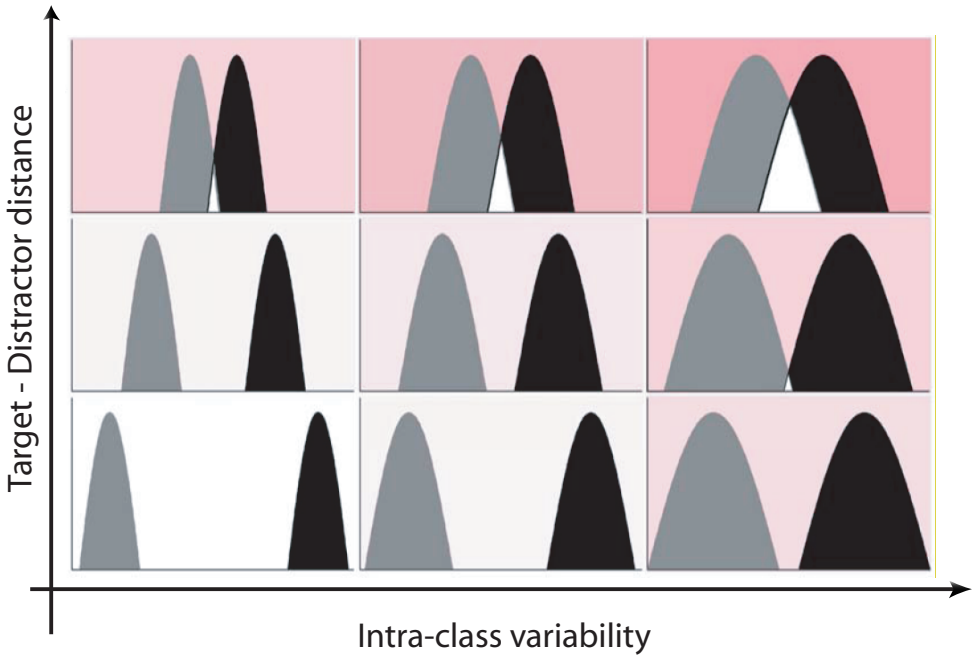


Figure 4.7: Schematic view of the impact of intra-class variability and distance between targets and distractors on the classification task difficulty (modified from (Macé, 2006)). Our results, in line with previous experiments (Macé, 2006), suggest that the second factor is more important than the first one in ultra-rapid visual categorization.

4.6.4 The gap shifts the speed-accuracy trade-off

Given the difficulty of the task, to reach a decent level of accuracy we had to remove the 200 ms gap that Kirchner and Thorpe (2006) used, at the expense of longer reaction times.

Generally speaking such a gap could provide an interesting leverage to move along the speed-accuracy curve. A negative gap, that is leaving the fixation cross for a few ms after the images have been flashed probably leads to even slower but more accurate responses. An adaptive gap could thus enable to normalize either the speed or the accuracy across subjects and/or conditions.

We are planning new experiments to test these predictions.

4.7 Conclusion

To conclude, this study is really just a first step – the 25 ms reduction in reaction time is indeed consistent with a decrease in latency to repeated stimuli, that could be due to STDP. However, many questions remain open, such as if this 25 ms speed-up is shift-invariant and depends on the type of stimuli, and if the difficulty encountered here comes from the nature of the task (intra-class identification) or from the image category (*e.g.* how fast and accurate would we be at identifying a given face picture among other face pictures?).

Given the variability of the behavioral reaction times, in the end, it may be that single unit recording studies will be the best level to look at this sort of latency decrease effect.

Chapter 5

Invariance learning: a plausibility proof

This work is the result of a collaboration with Thomas Serre and Tomaso Poggio, of the McGovern Institute for Brain Research, MIT, which started when I visited their laboratory for three months in summer 2006. We presented preliminary results at the *VSS 07* Conference, Sarasota, FL, USA (see Section B.3 for the conference abstract and poster), and we wrote a memo (Masquelier et al., 2007):

Masquelier T, Serre T, Thorpe S and Poggio T (2007) Learning complex cell invariance from natural videos: a plausibility proof. *CBCL Paper #269 / MIT-CSAIL-TR #2007-060, Massachusetts Institute of Technology, Cambridge, MA, USA*. `http://hdl.handle.net/1721.1/39833`

It is not in the 'main stream' of this thesis since rate coding is used. However it is complementary with respect to the studies presented in Chapters 2 and 3, which proposed mechanisms to learn *selectivity*: here we are interested in how *invariance* can be learnt. Future work will implement the proposed mechanism on spiking neurons.

5.1 Résumé

Une des caractéristiques les plus frappantes du cortex est sa capacité à s'auto-câbler. Comprendre comment ce câblage s'effectue durant le développement, et comment il s'affine ensuite par l'expérience tout au long de la vie est un des plus grands défis en neuroscience. Tandis que les modèles computationnels du système visuel deviennent de plus en plus détaillés (Riesenhuber

and Poggio, 1999; Giese and Poggio, 2003; Deco and Rolls, 2004, 2005; Serre et al., 2005a; Rolls and Stringer, 2006; Berzhanskaya et al., 2007; Masquelier and Thorpe, 2007), la question de comment la connectivité pourrait s'auto-organiser à partir de l'expérience visuelle est souvent négligée.

Ici on se focalise sur les modèles hiérarchiques de la voie ventrale du cortex visuel de type feedforward (Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997; LeCun and Bengio, 1998; VanRullen et al., 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Hochstein and Ahissar, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003; Serre et al., 2005a; Masquelier and Thorpe, 2007; Serre et al., 2007), qui étendent le modèle classique des cellules simples aux cellules complexes proposé par Hubel and Wiesel (1962) aux aires extra-striées, et dont on a montré qu'ils expliquent une quantité de données expérimentales. De tels modèles stipulent l'existence des deux sortes de cellules, les simples et les complexes, avec des prédictions spécifiques sur leurs connectivités respectives et les fonctionnalités qui en résultent.

Au sein de ces réseaux la question de l'apprentissage, particulièrement pour les cellules complexes, est probablement la moins bien comprise, et beaucoup d'auteurs ont recourt à des artifices tels que le câblage en dur et/ou le partage des poids ('weight-sharing'). Plusieurs algorithmes ont été proposés pour l'apprentissage des cellules complexes basé sur une 'trace rule' qui exploite la continuité temporelle du monde (ex (Földiák, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Spratling, 2005)), mais très peu sont capables d'apprendre à partir de séquences d'images naturelles non segmentées.

Nous proposons ici une nouvelle variante de la 'trace rule' qui renforce uniquement les synapses entre les cellules les plus activées, ce qui lui permet de fonctionner dans des environnements chargés. Jusqu'ici l'algorithme a été testé au niveau des cellules simples et complexes du cortex visuel primaire (V1). On a vérifié qu'une sélectivité de type Gabor pouvait émerger à partir d'un mécanisme d'apprentissage de type hebbien compétitif – ce qui avait été montré précédemment (ex (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006) – puis nous montrons comment la 'trace rule' modifiée permet aux cellules complexes situées en aval de se connecter à des cellules simples dont l'orientation préférée est la même, mais dont les champs récepteurs sont décalés.

5.2 Abstract

One of the most striking feature of the cortex is its ability to wire itself. Understanding how the visual cortex wires up through development and how visual experience refines connections into adulthood is a key question for Neuroscience. While computational models of the visual cortex are becoming increasingly detailed (Riesenhuber and Poggio, 1999; Giese and Poggio, 2003; Deco and Rolls, 2004, 2005; Serre et al., 2005a; Rolls and Stringer, 2006; Berzhanskaya et al., 2007; Masquelier and Thorpe, 2007), the question of how such architecture could self-organize through visual experience is often overlooked.

Here we focus on the class of hierarchical feedforward models of the ventral stream of the visual cortex (Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997; LeCun and Bengio, 1998; VanRullen et al., 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Hochstein and Ahissar, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003; Serre et al., 2005a; Masquelier and Thorpe, 2007; Serre et al., 2007), which extend the classical simple-to-complex cells model by Hubel and Wiesel (1962) to extrastriate areas, and have been shown to account for a host of experimental data. Such models assume two functional classes of simple and complex cells with specific predictions about their respective wiring and resulting functionalities.

In these networks, the issue of learning, especially for complex cells, is perhaps the least well understood, and many authors use hard-wired connectivity and/or weight-sharing. Several algorithms have been proposed for complex cell learning based on a trace rule to exploit the temporal continuity of the world (for *e.g.* (Földiák, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Spratling, 2005), but very few can learn from natural cluttered image sequences.

Here we propose a new variant of the trace rule that only reinforces the synapses between the most active cells, and therefore can handle cluttered environments. The algorithm has so far been developed and tested through the level of V1-like simple and complex cells: we verified that Gabor-like simple cell selectivity could emerge from competitive hebbian learning, as had been shown before (*e.g.* (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006)), and we show how the modified trace rule allow the subsequent complex cells to pool over simple cells with the same preferred orientation, but with shifted receptive fields.

5.3 Introduction

One of the most striking feature of the cortex is its ability to wire itself (see Section 1.1). Understanding how the visual cortex wires up through development and how plasticity refines connections into adulthood is likely to give necessary constraints to computational models of visual processing. Surprisingly there have been relatively few computational studies (Perrett et al., 1984; Földiák, 1991; Hietanen et al., 1992; Wallis et al., 1993; Wachsmuth et al., 1994; Wallis and Rolls, 1997; Stringer and Rolls, 2000; Rolls and Milward, 2000; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Spratling, 2005) that have tried to address the mechanisms by which learning and plasticity may shape the receptive fields of neurons in the visual cortex. Perhaps this can, in part, be explained by a relative lack of experimental data (see Section 1.3). At the same time, this is somehow a paradox as theoretical work could indeed provide interesting predictions to be tested experimentally.

Here we study biologically plausible mechanisms for the learning of selectivity and invariance of cells in the primary visual cortex (V1). We focus on a specific class of models of the ventral stream of the visual cortex, the feedforward hierarchical models of visual processing (Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997; LeCun and Bengio, 1998; VanRullen et al., 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Hochstein and Ahissar, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003; Serre et al., 2005a; Masquelier and Thorpe, 2007; Serre et al., 2007), which extend the classical simple-to-complex cells model by Hubel and Wiesel (1962) (see Box 1) and have been shown to account for a host of experimental data.

To be precise we have used the HMAX model (Riesenhuber and Poggio, 1999; Serre et al., 2005b, 2007), which assumes two functional classes of simple and complex cells with specific predictions about their respective wiring, and focused on the learning of the V1 simple and complex cells, respectively called S_1 and C_1. Learning in higher stages of the model will be addressed in future work. We show that with simple biologically plausible learning rules, these two classes of cells can be learned from natural real-world videos with no supervision. In particular, we verified that S_1 Gabor-like selectivity could emerge from competitive Hebbian learning, as had been showed before (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006), and, more importantly, we proposed a new mechanism, which suggests how the specific pooling from S_1 to C_1 could self-organize by passive exposure to natural input video sequences. We discuss the computational requirements for such unsupervised learning to take place and make specific experimental predictions.

5.4 HMAX Model

The key computational issue in object recognition is the specificity-invariance trade-off: recognition must be able to finely discriminate between different objects or object classes while at the same time be tolerant to object transformations such as scaling, translation, illumination, changes in viewpoint, changes in context and clutter, as well as non-rigid transformations (such as a change of facial expression) and, for the case of categorization, also to variations in shape within a class. Thus the main computational difficulty of object recognition is achieving a trade-off between selectivity and invariance. Extending the hierarchical model by (Hubel and Wiesel, 1959) (see Box 1) to extrastriate areas and based on theoretical considerations, Riesenhuber and Poggio (1999) speculated that only two functional classes of units may be necessary to achieve this trade-off, and demonstrated it with the so-called HMAX model[1]:

5.4.1 The *Simple* S units

They perform a TUNING operation over their afferents to build object-selectivity (the analog of the TUNING operation in computer vision is the *template matching* operation between an input image and a stored representation). The simple S units receive convergent inputs from retinotopically organized units tuned to *different preferred stimuli* and combine these *subunits* with a bell-shaped tuning function, thus increasing object selectivity and the complexity of the preferred stimulus (see (Serre et al., 2005a) for details).

As discussed in (Poggio and Bizzi, 2004) neurons with a Gaussian-like bell-shape tuning are prevalent across cortex. For instance simple cells in V1 exhibit a Gaussian tuning around their preferred orientation (Hubel and Wiesel, 1962) or even cells in inferotemporal cortex are typically tuned around a particular view of their preferred object (Logothetis et al., 1995; Booth and Rolls, 1998). From the computational point of view, Gaussian-like tuning profiles may be key in the generalization ability of cortex and networks that combine the activity of several units tuned with a Gaussian profile to different training examples have proved to be powerful learning scheme (Poggio and Girosi, 1990; Poggio and Smale, 2003).

5.4.2 The *Complex* C units

They receive convergent inputs from retinotopically organized S units tuned to the *same preferred stimuli* but at slightly different positions and scales with

[1]HMAX stands for Hierarchical MAXimum

Box 1: The Hubel & Wiesel hierarchical model of V1.

Following their work on striate cortex, Hubel & Wiesel described a hierarchy of cells in the primary visual cortex: At the bottom of the hierarchy, the *radially symmetric* cells are like LGN cells and respond best to small spots of light. Second, the *simple* cells do not respond well to spots of light and require bar-like (or edge-like) stimuli at a particular orientation, position and phase (*i.e.* white bar on a black background or dark bar on a white background). In turn, the *complex* cells are also selective for bars at a particular orientation but they are insensitive to both the location and the phase of the bar within their receptive fields. At the top of the hierarchy the *hypercomplex* cells not only respond to bars in a position and phase invariant way, just like complex cells, but are also selective for bars of a particular length (beyond a certain length their response starts decreasing).

Hubel & Wiesel suggested that such increasingly complex and invariant object representations could be progressively built by integrating convergent inputs from lower levels. For instance, as illustrated in Fig. 5.1 (reproduced from (Hubel and Wiesel, 1959)), position invariance at the complex cells level, could be obtained by pooling over simple cells at the same preferred orientation but at slightly different positions.

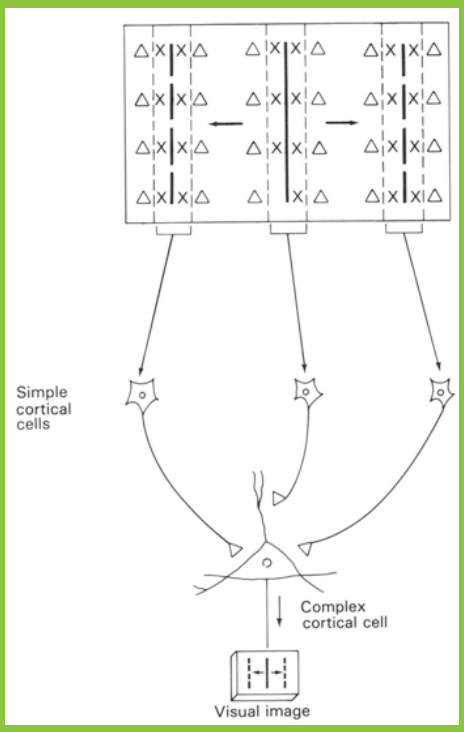


Figure 5.1: The Hubel & Wiesel hierarchical model for building complex cells from simple cells. Reproduced from (Hubel and Wiesel, 1959).

a MAX-like operation, thereby introducing tolerance to scale and translation. MAX functions are commonly used in signal processing (*e.g.* for selecting peak correlations) to filter noise out. The existence of a MAX operation in visual cortex was predicted by (Riesenhuber and Poggio, 1999) from theoretical arguments (and limited experimental evidence (Sato, 1989)) and was later supported experimentally in V4 (Gawne and Martin, 2002) and in V1 at the complex cell level (Lampl et al., 2004). Note that other recordings are more compatible with an average operation than with a MAX operation, see for example (Zoccolan et al., 2005).

5.4.3 Neural implementations of the two key operations

In this work we use static idealized approximation to describe the response of *populations* of simple and complex units (see Section 1.4.1 for a discussion on population coding). As described in (Serre et al., 2005a) both operations can be carried out by a *divisive normalization* followed by *weighted sum* and rectification. Normalization mechanisms (also commonly referred to as *gain control*) in this case, can be achieved by a feedforward (or recurrent) shunting inhibition (Torre and Poggio, 1978; Reichardt et al., 1983; Carandini and Heeger, 1994).[2]

The detailed mathematical formulation of the two operations is given in Box 2. There are plausible local circuits (Serre et al., 2005a) implementing the two key operations within the time constraints of the experimental data (Perrett et al., 1992; Keysers et al., 2001; Hung et al., 2005) based on small local *populations* of spiking neurons firing probabilistically in proportion to the underlying analog value (Smith and Lewicki, 2006) and on shunting inhibition (Grossberg, 1973). A complete description of the two operations, a summary of the evidence as well as plausible biophysical circuits to implement them can be found (Serre et al., 2005a) (see Section 5, pp. 53-59).

Other possibilities may involve spike timing in individual neurons (Masquelier and Thorpe, 2007) (see Section 1.5 for experimental evidence). Future

[2]For the past two decades several studies (in V1 for the most part) have provided evidence for the involvement of GABAergic circuits in shaping the response of neurons (Sillito, 1984; Douglas and Martin, 1991; Ferster and Miller, 2000). Direct evidence for the existence of divisive inhibition comes from an intracellular recording study in V1 (Borg-Graham and Fregnac, 1998). Wilson et al. (1994) also showed the existence of neighboring pairs of pyramidal cells / fast-spiking interneurons (presumably inhibitory) in the prefrontal cortex with inverted responses (*i.e.* phased excitatory/inhibitory responses). The pyramidal cell could provide the substrate for the weighted sum while the fast-spiking neuron would provide the normalization term.

Box 2: Computational implementation of the Hubel & Wiesel model.

We use a population coding framework (*i.e.* a 'unit' corresponds to a population of neurons). We denote by $(x_1, x_2, \ldots x_n)$ the set of inputs to a unit and $\mathbf{w} = (w_1, \ldots, w_N)$ the corresponding synaptic weights. For a complex unit, the inputs x_j are retinotopically organized and selected from an $m \times m$ grid of afferent units with the same selectivity (*e.g.* for an horizontal complex cells, subunits are all tuned to an horizontal bar but at slightly different positions and spatial frequencies). For a simple unit, the subunits are also retinotopically organized (selected from an $m \times m$ grid of possible afferents). But, in contrast with complex units, the subunits of a simple cell could in principal be with different selectivities to increase the complexity of the preferred stimulus.
Mathematically, both the TUNING operation and the MAX operation at the simple and complex units level can be well approximated by the following equation:

$$y = \frac{\sum_{j=1}^{n} w_j^* \, x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q \right)^r},$$

where y is the output of the unit, $k << 1$ is a constant to avoid zero-divisions and p, q and r represent the static non-linearities in the underlying neural circuit.
Such non-linearity may correspond to different regimes on the $f-I$ curve of the presynaptic neurons such that different operating ranges provide different degrees of non-linearities (from near-linearity to steep non-linearity). An extra sigmoid transfer function on the output $g(y) = 1/(1 + \exp^{\alpha(y-\beta)})$ controls the sharpness of the unit response.
By adjusting these non-linearities, the equation above can approximate better a MAX or a TUNING function:

- When $\mathbf{p} \lessapprox \mathbf{qr}$, the unit approximates a Gaussian-like TUNING, *i.e.* its response y will have a peak around some value proportional to the input vector $\mathbf{w} = (w_1, \ldots, w_N)$. For instance, when $p = 1$, $q = 2$ and $r = 1/2$, the circuits perform a normalized dot-product with an L_2 norm, which with the addition of a bias term may approximate a Gaussian function very closely (see (Kouh and Poggio, 2004; Serre et al., 2005a) for details).
- When $\mathbf{p} \gtrapprox \mathbf{q + 1}$ ($\mathbf{w_j \approx 1}$), the unit implements a soft-max and approximates a MAX function very closely for larger q values (see (Yu et al., 2002), the quality of the approximation also increases as the inputs become more dissimilar). For instance, $r \approx 1$, $p \approx 1$, $q \approx 2$ gives a good approximation of the MAX (see (Serre et al., 2005a) for details).

work will evaluate the use of spiking neuron and temporal coding in this framework.

While there exists at least partial evidence for the existence of both Gaussian TUNING and max-like operations (see earlier and (Serre et al., 2005a)), the question of how the specific wiring of simple and complex cells could self-organize during development and how their selectivity could be shape through visual experience is open. In the next section, we review related work and speculate on computational mechanisms that could underlie the development of such circuits.

5.5 On learning correlations

Here we speculate that correlations play a key role in learning. Beyond the Hebbian doctrine, which says that 'neurons that fire together wire together', we suggest that correlation in the inputs of neurons could explain the wiring of both simple and complex cells. As emphasized by several authors, statistical regularities in natural visual scenes may provide critical cues to the visual system to solve specific tasks (Richards et al., 1992; Knill and Richards, 1996; Callaway, 1998; Coppola et al., 1998) or even provide a teaching signal (Barlow, 1961; Sutton and Barto, 1981; Földiák, 1991) for learning with no supervision. More specifically, we suggest that the wiring of the simple S units depends on learning correlations in *space* while the wiring of the C units depends on learning correlations in *time*.

5.5.1 Simple cells learn spatial correlations

Simple cells learn spatial correlation between inputs *at the same time*(*i.e.* for simple S_1 units in V1, the bar-like arrangements of LGN inputs, and beyond V1, more elaborate arrangements of bar-like subunits, *etc.*). This corresponds to learning which combinations of features appear most frequently in images. That is, a simple unit has to detect *conjunctions* of inputs (*i.e.* sets of inputs that are consistently co-active), and to become selective to these patterns. This is roughly equivalent to learning a dictionary of image patterns that appear with higher probability.

This is a very simple and natural assumption. Indeed it follows a long tradition of researchers that have suggested that the visual system, through visual experience and evolution, may be adapted to the statistics of its natural environment (Attneave, 1954; Barlow, 1961; Atick, 1992; Ruderman, 1994) (see also (Simoncelli and Olshausen, 2001) for a review). For instance, (Attneave, 1954) proposed that the goal of the visual system is to build an

efficient representation of the visual world and (Barlow, 1961) emphasized that neurons in cortex try to reduce the redundancy present in the natural environment.

This type of learning can be done with an Hebbian learning rule (Földiák, 1990). Here we used a slightly modified Hebb rule, which has the advantage of keeping the synaptic weights bounded, while remaining a local learning rule (see Equation 5.6). At the same time, a mechanism is necessary to prevent all the simple units in a given cortical column from learning the same pattern. Here we used hard competition of the 1-Winner-Take-All form (see (Rolls and Deco, 2002) for evidence). In the algorithm we describe below, at each iteration and within each hypercolumn only the most activated unit is allowed to learn (but it will do so if and only if its activity is above a threshold, see Section 5.8.3). In the cortex such a mechanism could be implemented by short range lateral inhibition.

Networks with anti-Hebbian horizontal connections have also been proposed (Földiák, 1990). While such networks could, in principle, remove redundancy more efficiently, they cannot account for the initial feedforward response of neurons within the first 10-30 ms after response onset (Thorpe and Imbert, 1989; Thorpe and Fabre-Thorpe, 2001; Keysers et al., 2001; Rolls, 2004) but could nevertheless be easily added in future work. Furthermore, a certain level of redundancy is desirable, to handle noise and loss of neurons.

Matching pursuit, that could also be implemented in visual cortex through horizontal connections, has also been proposed to reduce the redundancy and increases the sparseness of neuronal responses (Perrinet et al., 2004a).

Previous work has already shown how selectivity to orientation could emerge naturally with simple learning rules like Spike-Timing-Dependant-Plasticity (STDP) (Delorme et al., 2001; Guyonneau, 2006) and a hebbian rule (Einhäuser et al., 2002). The goal of the work here is to apply such rule to the specific HMAX model (Riesenhuber and Poggio, 1999; Serre et al., 2005a, 2007).

5.5.2 Complex cells learn temporal correlations

Complex cells may learn from visual experience how to associate frequent transformations in time – such as translation and scale – of specific visual features coded by simple cells. The wiring of the C units reflects learning of correlations *across time*, *e.g.* for complex C_1 units, learning which afferent S_1 units with the same orientation and neighboring locations should be wired together because, often, such a pattern changes smoothly in time (under translation) (Földiák, 1991; Wiskott and Sejnowski, 2002).

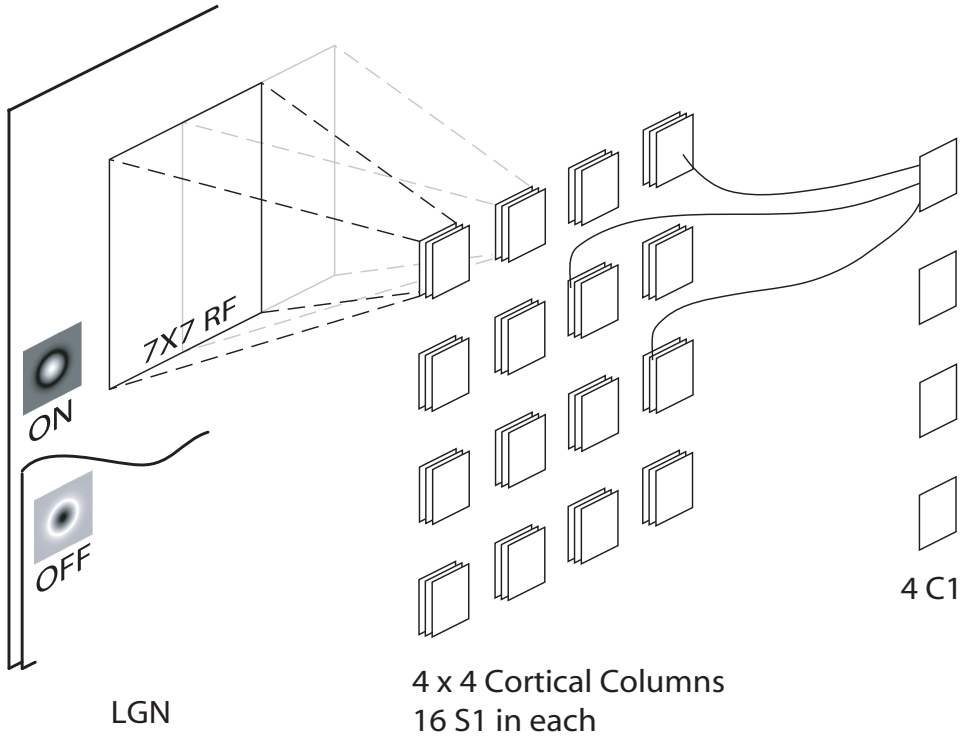


Figure 5.2: Overview of the specific implementation of the Hubel & Wiesel V1 model used. LGN-like ON- and OFF-center units are modeled by Difference-of-Gaussian (DoG) filters. Simple units (denoted S_1) sample their inputs from a 7×7 grid of LGN-type afferent units. Simple S_1 units are organized in cortical hypercolumns (4×4 grid, 3 pixels apart, 16 S_1 units per hypercolumn). At the next stage, 4 complex units C_1 cells receive inputs from these $4 \times 4 \times 16$ S_1 cells. This chapter focuses on the learning of the S_1 to C_1 connectivity.

As discussed earlier, the goal of the complex units is to increase the invariance of the representation along one stimulus dimension. This is done by combining the activity of a group of neighboring simple units tuned to the same preferred stimulus at slightly different position and scales. In this chapter we focus on translation invariance, but the same mechanism could be applied to any transformation (for *e.g.* scale, rotation or view-point).

A key question is how a complex cell would 'know' which simple cells it should connect to, *i.e.* which simple cells do represent the same object at different locations? Note that a standard Hebbian rule, that learns conjunctions of inputs, does not work here, as only one (or a few) of the targeted simple cells will be activated at once. Instead, a learning rule is needed to learn *disjunctions* of inputs.

Several authors have proposed to use temporal continuity to learn complex cells from transformation sequences (Perrett et al., 1984; Földiák, 1991; Hietanen et al., 1992; Wallis et al., 1993; Wachsmuth et al., 1994; Wallis and Rolls, 1997; Rolls and Milward, 2000; Wiskott and Sejnowski, 2002). This can be done using an associative learning rules that incorporate a temporal trace of activity in the post-synaptic neuron (Földiák, 1991), exploiting the fact that objects seldom appear or disappear, but are often translated in the visual field. Hence simple units that are activated in close temporal proximity are likely to represent the same object, presumably at different locations. Földiák (1991) proposed a modified Hebbian rule, known as the 'trace rule' which constrain synapses to be reinforced when strong inputs coincides with strong average past activity (instead of strong current activity in case of a standard Hebbian rule). This proposal has formed the basis of a large number of algorithms for learning invariances from sequences of images (Becker and Hinton, 1992; Stone and Bray, 1995; Wallis and Rolls, 1997; Bartlett and Sejnowski, 1998; Stringer and Rolls, 2000; Rolls and Milward, 2000; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Spratling, 2005).

However, as pointed out by Spratling (2005), the trace rule by itself is inappropriate when multiple objects are present in a scene: it cannot distinguish which input corresponds to which object, and it may end-up combining multiple objects in the same representation. Hence most trace-rule based algorithm require stimuli to be presented in isolation ((Földiák, 1991; Oram and Földiák, 1996; Wallis, 1996; Stringer and Rolls, 2000)), and would fail to learn from cluttered natural input sequences.

To solve this problem, Spratling made the hypothesis that the same object could not activate two distinct inputs, hence co-active units necessarily correspond to distinct objects. He proposed a learning rule that can exploit this information, and successfully applied it on drifting bar sequences(Spratling, 2005).

However the 'one object activates one input' hypothesis is a strong one. It seems incompatible with the redundancy observed in the mammalian brain and reproduced in our model. Instead we propose another hypothesis: from one frame to another the most active inputs are likely to represent the same object. If the hypothesis is true, by restraining the reinforcement to the most active inputs we usually avoid to combine different objects in the same representation (note that this idea was already present in (Einhäuser et al., 2002), although not formulated in those terms).

In this chapter we focus on the learning of simple S_1 and complex C_1 units (see Fig. 5.2), which constitutes a direct implementation of the Hubel and Wiesel (1959) model of striate cortex (see Box 1). The goal of a C_1 unit is to pool over S_1 units with the same preferred orientation, but with shifted receptive fields. In this context our hypothesis becomes: 'in a given neighborhood, the dominant orientation is likely to be the same from one frame to another'. As our experimental suggests, this constitutes a reasonable hypothesis, which leads to appropriate pooling.

5.6 Results

We tested the proposed learning mechanisms in a 3 layer feedforward network mimicking the Lateral Geniculate Nucleus (LGN) and V1 (see Fig. 5.2). Details of the implementation can be found in the Section 5.8.

The stimuli we used were provided by Betsch et al. (2004). The videos were captured by CCD cameras attached to cats' heads, while the animals were exploring several outdoor environments. Theses videos approximate the input to which the visual system is naturally exposed, although eye movements are not taken into account.

To simplify the computations, learning was done in two phases: First S_1 units learned their selectivity through competitive Hebbian learning. After convergence, plasticity at the S_1 stage was switched off and learning at the complex C_1 unit level started. In a more realistic scenario, this two-phase learning scheme could be approximated with a slow time constant for learning at the S_1 stage and a faster time constant at the C_1 stage.

5.6.1 Simple cells

After about 9 hours of simulated time S_1 units have learned a Gabor-like selectivity (see Fig. 5.3) similar to what has been previously reported for cortical cells (Hubel and Wiesel, 1959, 1962, 1965, 1968; Schiller et al., 1976a,b,c; DeValois et al., 1982a,b; Jones and Palmer, 1987; Ringach, 2002). In par-

ticular, receptive fields are localized, tuned to specific spatial frequencies in a given orientation. In this experiment, only four dominant orientations emerged spanning the full range of orientations with 45° increment: 0° , 45° , 90° and 135° . Interestingly, in an another experiment using S_1 receptive fields larger than the 7×7 receptive field sizes used here, we found instead a continuum of orientations (data not shown). The fact that we obtain only four orientations here is likely to be a discretization artifact. With this caveat in mind, in the following we used the 7×7 RF sizes (see Table 5.1), which match the receptive field sizes of cat LGN cells.

Our results are in line with previous studies that have shown that competitive Hebbian learning with DoG inputs leads to Gabor-like selectivity (see for instance (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006)).

5.6.2 Complex cells

In phase 2, to learn the receptive fields of the C_1 units, we turned off learning at the S_1 stage and began to learn the $S_1 - C_1$ connectivity. This was done using a learning rule that reinforce the synapse between the currently most activated S_1 unit and the previously most activated C_1 unit (see Section 5.8.4). After 19 hours of simulated time, we ended up with binary $S_1 - C_1$ weights, and each C_1 remained connected to a pool of S_1 with the same preferred orientation, eventually in different cortical columns (see Fig. 5.4). Hence by taking the (soft) maximum response among its pool, a C_1 unit becomes shift-invariant and inherits its orientation selectivity from its input S_1 units.

In total 38 S_1 units were not selected by any $C1$ (see Fig. 5.5). They either had an atypical preferred stimulus or were tuned to an horizontal bar, which, because of a possible bias in the training data (maybe due to horizontal head movements), is over-represented at the S_1 level. In addition we did not find any S_1 unit selected by more than one C_1 unit. In other words, the pools Fig. 5.4(a), 5.4(b), 5.4(c), 5.4(d) and 5.5 were all disjoint. Note that superimposing those 5 figures leads to Fig. 5.3.

We also experimented other learning rules in the same architecture (data not shown). We reimplemented Földiák's original trace rule (Földiák, 1991) given by:

$$\Delta \boldsymbol{w} = \alpha \cdot tr(y) \cdot (\boldsymbol{x} - \boldsymbol{w}). \tag{5.1}$$

As expected, the learning rule failed mainly due to the fact that input frames do not contain isolated edges but instead edges with multiple orientations. This, in turn, leads to complex units that pool over multiple orientations.

We also implemented Spratling's learning rule (Spratling, 2005), which

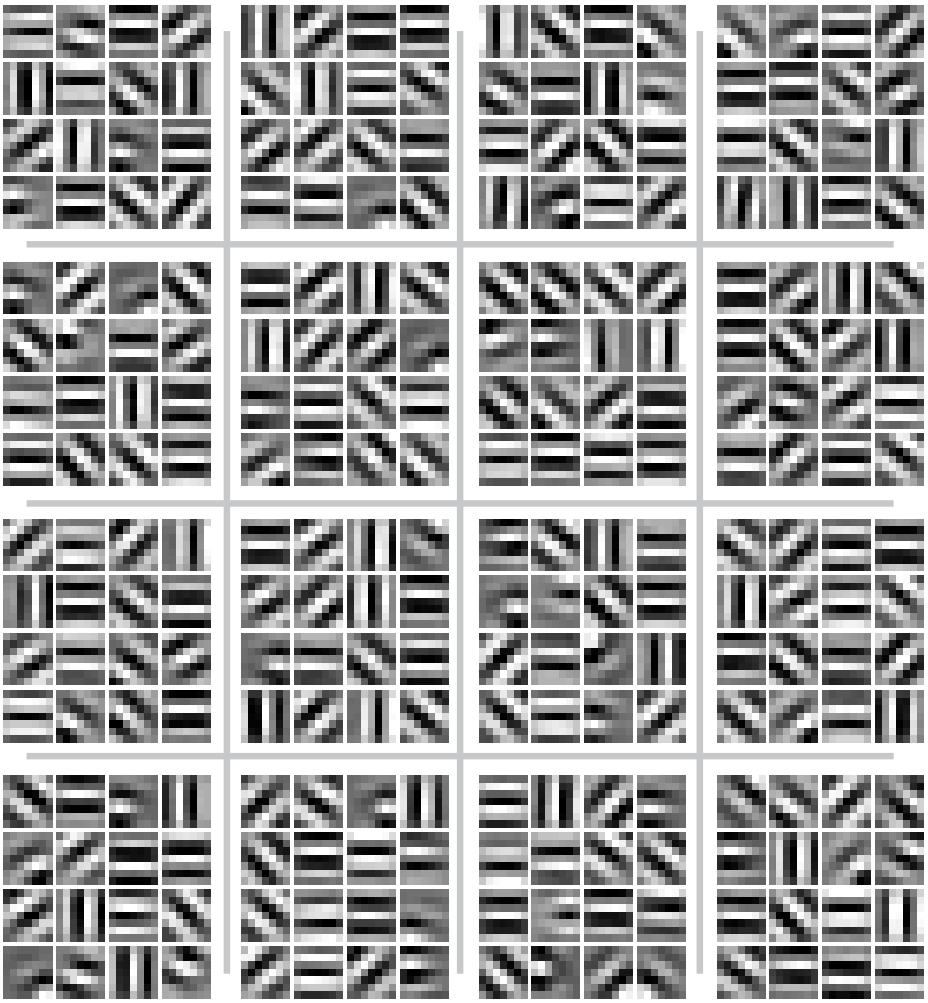


Figure 5.3: Reconstructed S_1 preferred stimuli for each one of the 4×4 cortical hypercolumns (on this figure the position of the reconstructions within a cortical column is arbitrary). Most units show a Gabor-like selectivity similar to what has been previously reported in the literature (see text).

failed in a similar way, because the hypothesis that 'one edge activates one S_1 unit' is violated here.

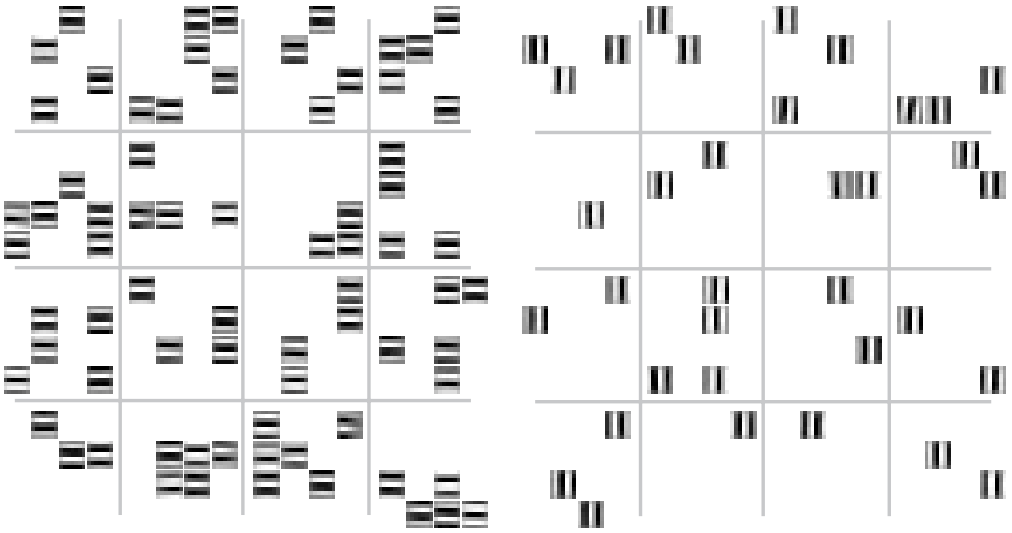
Finally we re-implemented the rule by Einhäuser et al. (2002) (see Equation 5.11). We reproduced their main results and the learning rule generated a continuum of $S_1 - C_1$ weights (as opposed to our binary weights). The strongest synapses of a given complex cell did correspond to simple cells with the same preferred orientation but, undesirably, the complex cells had also formed connection to other simple cells with distinct preferred orientation (see also Section 5.7).

5.7 Discussion

Contrary to most previous approaches (Földiák, 1990, 1991; Wallis and Rolls, 1997; Stringer and Rolls, 2000; Rolls and Milward, 2000; Spratling, 2005), our approach deals with natural image sequences, as opposed to artificial stimuli such as drifting bars. For a given algorithm to be biologically plausible a necessary condition (although not sufficient) is that it can handle natural images, which bring supplementary difficulties such as noise, clutter and absence of relevant stimuli. Models that process simpler stimuli may be useful to illustrate a given mechanism, but the goal *in fine* should be to deal with natural images, just like humans do. To our knowledge, the only model for the learning of simple and complex cells, which has been shown to work on natural image sequences is the one by Einhäuser et al. (2002). Our work extends the study by Einhäuser et al. (2002) in several significant ways.

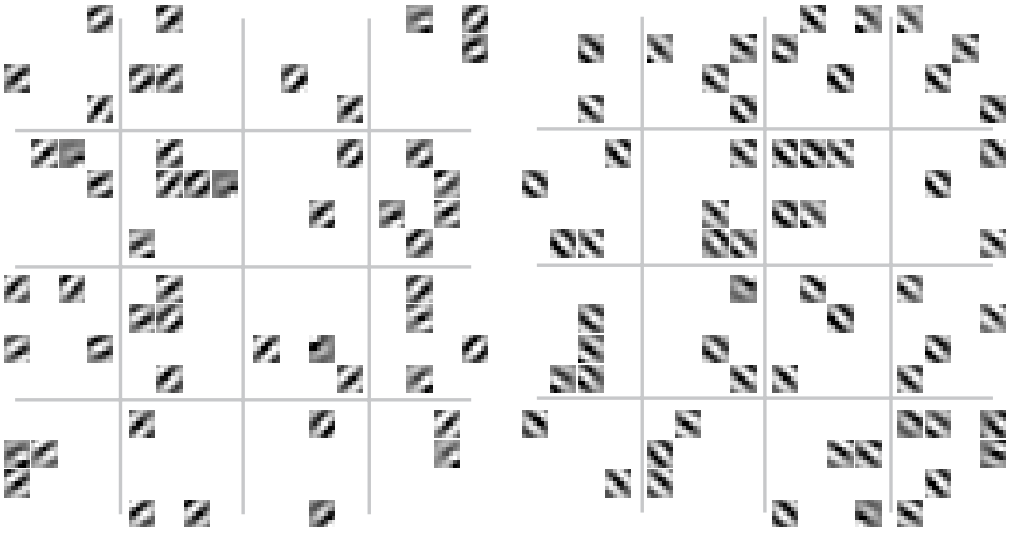
First, by using soft-bounds in the weight update rule (see Equation 5.10), the proposed algorithm converges towards input weights to a complex unit that are binary. This means that each complex unit is strongly connected to a pool of simple units that all have the same preferred orientation. This leads to complex units with an orientation bandwidth similar to the orientation bandwidth of simple units (see (Serre and Riesenhuber, 2004)) in agreement with experimental data (DeValois et al., 1982b). Conversely, the algorithm by Einhäuser *et al.* generates a continuum of synaptic weights and although weaker, some of the connections to simple units to non-preferred orientations remained thus broadening the orientation bandwidth from simple to complex units.

Another important difference with the learning rule used in (Einhäuser et al., 2002) is that our modified Hebbian learning rule is based on the correlation between the current inputs to a complex unit and its output at the previous time step (as opposed to previous input and current output in (Einhäuser et al., 2002)). This was suggested in (Rolls and Milward, 2000). Here



(a) S_1 units (n=73) that remain connected to C_1 unit # 1 after learning

(b) S_1 units (n=35) that remain connected to C_1 unit # 2 after learning



(c) S_1 units (n=59) that remain connected to C_1 unit # 3 after learning

(d) S_1 units (n=38) that remain connected to C_1 unit # 4 after learning

Figure 5.4: Pools of S_1 units connected to each C_1 unit. For *e.g.* C_1 unit # 1 became selective for horizontal bars: After learning only 73 S_1 units (out of 256) remain connected to the C_1 unit, and they are all tuned to an horizontal bar, but at different positions (corresponding to different cortical columns; on this figure the positions of the reconstructions correspond to their positions in Fig. 5.3).

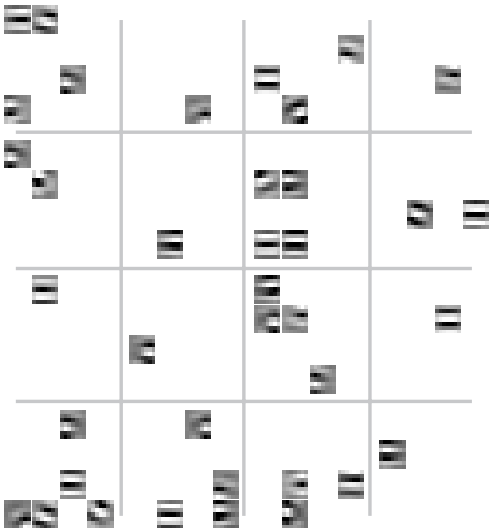


Figure 5.5: The 38 S_1 cells that were not connected to any C_1.

we found empirically that it leads to faster and more robust learning. It also turns out to be easier to implement in biophysical circuits: Because of synaptic delays, it is in fact very natural to consider the current input to a unit and its output to the previous frames a few tens of milliseconds earlier. Measuring correlations between past inputs and current output would need an additional mechanism to store the current input for future use.

Finally our approach tends to be simpler than most of the previous ones. The inputs to the model are raw gray-level images without any pre-processing such as low pass filtering or whitening. Also the proposed algorithm does not require any weight normalization and all the learning rules used are local.

Our neurophysiologically-plausible approach also contrasts with objective function approaches, which optimize a given function (such as sparseness (Olshausen and Field, 1996; Rehn and Sommer, 2007) (minimizing the number of units active for any input), statistical independence (Bell and Sejnowski, 1997; van Hateren and Ruderman, 1998; van Hateren and van der Schaaf, 1998; Hyvärinen and Hoyer, 2001) or even temporal continuity and slowness (Wiskott and Sejnowski, 2002; Körding et al., 2004; Berkes and Wiskott, 2005) in a non-biologically plausible way. Such normative model can provide insights as to why receptive fields look the way they do. Indeed such models have made quantitative predictions, which have been compared to neural data (see (van Hateren and Ruderman, 1998; van Hateren and van der Schaaf, 1998; Ringach, 2002) for instance). However, such approaches ignore

the computational constraints imposed by the environment and are agnostic about how such learning could be implemented in the cortex.

Fortunately connections can be drawn between the two classes of approaches. For example, Sprekler *et al*. recently showed that Slow Feature Analysis (SFA) is in fact equivalent to the trace rule, and could be implemented by Spike Timing Dependant Plasticity (Sprekeler et al., 2007).

Finally our approach constitutes a plausibility proofs for most models of the visual cortex (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2007; Masquelier and Thorpe, 2007), which uses weight-sharing (see Section 1.6.3). Such model typically learn at one location and simply 'replicate' units at all locations. This is not the approach we undertook in this work: The 4×4 grid of S_1 units (16 at each location) are all learned independently and indeed are not identical. We then suggested a mechanism to pool together cells with similar preferred stimulus. The success of our approach validates the simplifying assumption of weight-sharing. However we still have to test the proposed mechanisms for higher order neurons (such as the C2 cells of Chapter 2).

The idea of exploiting temporal continuity to build invariant representations finds partial support from psychophysical studies, which have suggested that human observers tend to associate together successively presented views of paperclip objects (Sinha and Poggio, 1996) or faces (Wallis and Bülthoff, 2001). The idea also seems consistent – as pointed out by Stryker (Stryker, 1991; Földiák, 1998; Giese and Poggio, 2003) – with an electrophysiological study by Miyashita (1988), who showed, that training a monkey with a fixed sequence of image patterns lead to a correlated activity between those same patterns during the delayed activity.

Finally this class of algorithms lead to an interesting prediction made by Einhäuser et al. (2002), namely that the selectivity of complex units could be impaired by rearing an animal in an environment in which temporal continuity would be disrupted (for instance using a stroboscopic light or constantly flashing uncorrelated pictures). We verified this prediction on our model and found that randomly shuffling the frames of the videos had no impact on the development of the simple S_1 units while the selectivity of the complex C_1 units was significantly impaired (all the synapses between the simple and complex units ended up depressed).

To conclude, although this study could be pushed further – in particular the proposed mechanism should be implemented on spiking neurons, and should be tested on higher order neurons – it does constitute au plausibility proof that invariances could be learnt using a simple trace-rule, even in natural cluttered environment.

Figure 5.6: Videos: the world from a cat's pespective (Betsch et al., 2004).

5.8 Technical details

5.8.1 Stimuli: the world from a cat's perspective

The videos used were taken from (Betsch et al., 2004)(see Fig. 5.6). The camera spans a visual angle of 71° by 53° and its resolution is 320×240 pixels. Hence each pixel corresponds to about 13 min of arc. We only used the first six videos (from thirteen total) for a total duration of about 11 minutes. Spatio-temporal patches were extracted from these videos at fixed points from a 9×11 grid (sampled every 25 pixels). These 99 sequences were concatenated leading to a total of about 19 hours of video (about 1.6 million frames).

In the following, we set the receptive field sizes for model LGN-like, simple S_1 and complex C_1 units to the average values reported in the literature for foveal cells in the cat visual cortex (Hubel and Wiesel, 1968). We did not model the increase in RF size with eccentricity and assumed that foveal values stood everywhere. This leads to receptive field sizes for the three layers that are summarized in Table 5.1.

Table 5.1: Receptive field sizes in pixels, and in degree of visual angle.

	$ON-OFF$	S_1	C_1
Pixels	7	13	22
Degrees	1.6	2.9	4.9

5.8.2 LGN ON- and OFF-center unit layer

Gray level images are first analyzed by an array of LGN-like units that correspond to 7×7 Difference-of-Gaussian (DoG) filters:

$$DoG = \frac{1}{2\pi}\left(\frac{1}{\sigma_1}\, e^{-\frac{r^2}{2\sigma_1^2}} - \frac{1}{\sigma_2}\, e^{-\frac{r^2}{2\sigma_2^2}}\right) \tag{5.2}$$

We used $\sigma_2 = 1.4$ and $\sigma_2/\sigma_1 = 1.6$ to make the DOG receptive fields approximate a Laplacian filter profile, which in turn resembles the receptive fields of biological retinal ganglion cells (Marr and Hildreth, 1980). Positive values ended in the ON-center cell map, and the absolute value of negative values in the OFF-center cell map.

5.8.3 S_1 layer: competitive hebbian learning

Model S_1 units are organized on a 4×4 grid of cortical columns. Each column contains 16 S_1 units (see Fig. 5.2). The distance between columns was set to 3 pixels (*i.e.* about half a degree of visual angle). Each S_1 unit received their inputs from a 7×7 grid of afferent LGN-like units (both ON- and OFF-center) for a total of $7 \times 7 \times 2$ input units. S_1 units perform a bell-shape TUNING (see (Serre et al., 2005a) for details) function which can be approximated by the following static mathematical operation (see Box 2):

$$\boldsymbol{y}_{\text{raw}} = \frac{\sum_{j=1}^{n} w_j.x_j^p}{k + (\sum_{j=1}^{n} x_j^q)^r} \tag{5.3}$$

Here for the S_1 cells we set the parameters to: $k = 0$, $p = 1$, $q = 2$ and $r = 1/2$, which is exactly a normalized dot-product:

$$y_{\text{raw}} = \frac{\boldsymbol{w}.\boldsymbol{x}}{||\boldsymbol{x}||} \tag{5.4}$$

The reader should refer to (Knoblich et al., 2007) for biophysical circuits of integrate and fire neurons that use realistic parameters of synaptic

transmission approximating Equation 5.4.

The response of a simple S_1 unit is maximal if the input vector x is collinear to the synaptic weight vector w (*i.e.* the preferred stimulus of the unit). As the pattern of input becomes more dissimilar to the preferred stimulus, the response of the unit decreases monotonically in a bell-shape-like way (*i.e.* the cosine of the angle between the two vectors).

The unit activity y_{raw} is further normalized by the recent unit history, *i.e.* a 'running average (denoted by $tr(.)$) of the raw activities over past few frames':

$$y = \frac{y_{\text{raw}}}{tr(y_{\text{raw}})} \tag{5.5}$$

Such unit history is often referred to as a (memory) *trace* (Földiák, 1991; Wallis, 1996; Wallis and Rolls, 1997; Stringer and Rolls, 2000; Rolls and Milward, 2000). For our model S_1 unit, such normalization by the trace approximates adaptation effects. One can think of y_{raw} as the membrane potential of the unit while y approximates the instantaneous firing rate of the unit over short time intervals: units that have been strongly active will become less responsive. While non-critical, this normalization by the trace significantly speeds-up the convergence of the learning algorithm by balancing the activity between all S_1 units (the response of units with a record of high recent activity is reduced while the response of units which have not been active in the recent past is enhanced).[3]

The initial $\boldsymbol{w}$ weights of all the S_1 units were initialized at random (sampled from a uniform distribution on the [0,1] interval). In each cortical column only the most active cell is allowed to fire (1-Winner-Take-All mechanism). However, it will do so if and only if its activity reaches its threshold T. It will then trigger the (modified) Hebbian rule:

$$\Delta \boldsymbol{w} = \alpha \cdot y \cdot (\boldsymbol{x} - \boldsymbol{w}) \tag{5.6}$$

The $-w$, added to the standard Hebb rule, allows to keep the w bounded. However, the learning rule is still fully local.

The winner then updates its threshold as follows:

$$T = y \tag{5.7}$$

[3]Empirically we found that the learning algorithm would still converge without the normalization term. However the distribution of preferred orientations among the learned S_1 units would be far less balanced than in the full learning algorithm (the number of horizontal units would outweigh the number of vertical ones (data not shown).

At each time step, all thresholds are decreased as follows:

$$T = (1 - \eta) \cdot T \tag{5.8}$$

There is experimental evidence for such threshold modulations in pyramidal neurons, which contribute to homeostatic regulation of firing rates. For example, Desai *et al.* showed that depriving neurons of activity for two days increased sensitivity to current injection (Desai et al., 1999).

At each time step the traces are updated as follows:

$$tr(y_{\text{raw}}) = \frac{y_{\text{raw}}}{\nu} + (1 - \frac{1}{\nu}) \cdot tr(y_{\text{raw}}) \tag{5.9}$$

We used $\eta = 2^{-15}$ and $\nu = 100$. It was found useful to geometrically increase the learning rate α for each S_1 cell every 10 weight updates, starting from an initial value of 0.01 and ending at 0.1 after 200 weight updates. Only half of the 1,683,891 frames were needed to reach convergence.

5.8.4 C_1 Layer: pool together consecutive winners

4 C_1 cells receive inputs from the $4 \times 4 \times 16$ S_1 cells through synapses with weight $w \in [0, 1]$ (initially set to .75).

Each C_1 cell's activity is computed using equation 5.3, but this time with $p = 6$ (and still $q = 2$ and $r = 1/2$). It has been shown that such operation performs a SOFT-MAX (Yu et al., 2002), and biophysical circuits to implement it have been proposed in (Knoblich et al., 2007).

Winner-Take-All mechanisms select the C_1 winner at time $t-\Delta t$ (previous frame), $J_{t-\Delta t}$, and the current S_1 winner at time t (current frame), I_t. The synapse between them is reinforced, while all the other synapses of $J_{t-\Delta t}$ are depressed:

$$\Delta w_{iJ_{t-\Delta t}} = \begin{cases} a^+ \cdot w_{iJ_{t-\Delta t}} \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{if } i = I_t \\ a^- \cdot w_{iJ_{t-\Delta t}} \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{otherwise} \end{cases} \tag{5.10}$$

Synaptic weights for the non-winning C_1 cells are unchanged.

This learning rule was inspired by previous work on Spike Timing Dependent Plasticity (STDP) (Masquelier and Thorpe, 2007). The multiplicative term $w_{iJ_{t-\Delta t}} \cdot (1 - w_{iJ_{t-\Delta t}})$ ensures the weight remains in the range [0,1] (excitatory synapses) and implements a soft bound effect: when the weight approaches a bound, weight changes tend toward zero, while the most plastic synapses are those in an intermediate state.

As recommended by (Rolls and Milward, 2000) we chose to exploit correlations between the previous output and the current input (as opposed to

current output and previous output, as in Einhäuser *et al.*'s model (Einhäuser et al., 2002)). We empirically confirmed that learning was indeed more robust this way.

It was found useful to geometrically increase the learning rates every 1000 iteration, while maintaining the a^+/a^- ratio at a constant value (-170). We started with $a^+ = 2^{-3}$ and set the increase factor so as to reach $a^+ = 2^{-1}$ at the end of the simulation.

5.8.5 Main differences with Einhäuser *et al.* 2002

- Learning rule for complex cells: first Einhäuser *et al.* select the C_1 winner at time t (current frame), J_t, and the previous S_1 winner at time $t - \Delta t$ (previous frame), $I_{t-\Delta t}$. The synapse between them is reinforced, while all the other synapses of J_t are depressed. Second, they use a different weight update rule:

$$\Delta w_{iJ_t} = \begin{cases} \alpha \cdot (1 - w_{iJ_t}) & \text{if } i = I_{t-\Delta t} \\ -\alpha \cdot w_{iJ_t} & \text{otherwise} \end{cases} \tag{5.11}$$

  This learning rule leads to a continuum of weights at the end (as opposed to binary weights) Tests have shown that the problem persists if (like us) we select the C_1 winner at time $t - \Delta t$ (previous frame), $J_{t-\Delta t}$, and the current S_1 winner at time t (current frame), I_t, and apply Einhäuser's rule:

$$\Delta w_{iJ_{t-\Delta t}} = \begin{cases} \alpha \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{if } i = I_t \\ -\alpha \cdot w_{iJ_{t-\Delta t}} & \text{otherwise} \end{cases} \tag{5.12}$$

  so the problem comes from the weight update rule they use, and not from the type of correlation involved.

- They normalize activity by running averages also at the complex stage
- They did not model the increase in RF size between S_1 and C_1

Chapter 6

Conclusions

6.1 Résumé

En introduction j'ai décrit trois propriétés essentielles des réponses neuronales dans le cortex visuel. Elles sont :

1. Sélectives (voir Section 1.2.1, paragraph 1)
2. Invariantes (voir Section 1.2.1, paragraph 2)
3. Rapides (voir Section 1.2.2)

Au Chapitre 2 j'ai proposé un mécanisme basé sur la STDP qui pourrait expliquer à la fois la sélectivité et la vitesse des réponses. Plus précisément, j'ai démontré que, au sein d'un model neuronal de la voie ventrale de type 'feed-forward', la combinaison d'une part d'un schéma de codage temporel dans lequel les neurones les plus stimulés déchargent en premier, et d'autre part de la STDP, amène à une situation dans laquelle les neurones des aires de haut niveau deviennent graduellement sélectifs à des combinaisons fréquentes de primitives visuelles. En outre, les réponses de ces neurones deviennent de plus en plus rapides. Le modèle est attrayant parce que, comme d'autres réseaux hiérarchiques multicouches (du type Fukushima (1980); LeCun and Bengio (1998); Riesenhuber and Poggio (1999); Wallis and Rolls (1997); Rolls and Milward (2000); Stringer and Rolls (2000); Serre et al. (2007)), il permet une reconnaissance d'objet robuste tout en évitant une explosion combinatoire, mais aussi parce que la reconnaissance est rapide, comme suggéré par la littérature expérimentale (Oram and Perrett, 1992; Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Keysers et al., 2001; Rousselet et al., 2002; Bacon-Mace et al., 2005; Hung et al., 2005; Kirchner and Thorpe, 2006; Serre et al., 2007; Girard et al., 2007).

Une prédiction intéressante de ce modèle est que les latences des réponses visuelles devraient diminuer après présentations répétées d'un même stimu-

lus. Au Chapitre 4 j'ai testé expérimentalement cette prédiction, en inférant les temps de traitement visuels à partir de mesures comportementales. J'ai pu confirmer que la familiarité avec une image peut diminuer le temps nécessaire pour la reconnaître d'environ 25 ms, et ce après quelques centaines de présentation seulement. Bien sûr cela ne veut pas dire que le modèle est vrai – seulement qu'il est plausible.

Un mécanisme basé sur une nouvelle variante de la 'trace rule' qui pourrait expliquer certaines invariances des réponses a été proposé au Chapitre 5. Certes cette étude pourrait être complétée (en particulier les mécanismes proposés devraient être implémentés sur des neurones impulsionnels, puis testés sur des couches de plus haut niveau), mais elle montre comment, même dans des environnements chargés, des neurones pourraient exploiter la continuité temporelle du monde pour construire des représentations invariantes.

Même si les modèles proposés sont encore un peu spéculatifs, ils s'appuient sur des mécanismes biophysiques simples et communément admis. Ils sont donc biologiquement plausibles.

Au delà de ces résultats propres au système visuel les travaux présentés ici créditent également l'hypothèse de l'utilisation de codage temporel dans le cerveau. En particulier l'étude présentée au Chapitre 3 est très générique : elle n'est pas restreinte à la vision ni même aux systèmes sensoriels. On a montré que, étonnamment, la STDP permet de détecter des patterns de spikes spatio-temporels même s'ils sont insérés dans des trains de spikes 'distracteurs' de même densité – un problème computationnellement complexe. La STDP permet donc l'utilisation d'un codage temporel, même en l'absence d'une date de référence explicite. Ce modèle et le modèle du système visuel présenté Chapitre 2 démontrent donc comment le cerveau pourrait facilement tirer profit de l'information contenue dans des dates de spikes. Si ces dates contiennent d'avantage d'information par rapport au taux de décharge moyen – la théorie dite du 'codage temporel' – est controversé. Etant donné que les mécanismes proposés ici sont à la fois simples, efficaces, et satisfont les contraintes temporelles provenant de la littérature expérimentale, ils constituent un argument fort en faveur du codage temporel.

Cela ne veut pas dire que le codage par taux de décharge n'est jamais utilisé, ni que les taux contiennent systématiquement moins d'information que les dates de spikes. Simplement, je pense que les schémas de codages temporels décrive mieux les régimes transitoires, qui sont probablement un aspect important de la computation neurale, surtout quand il s'agit de traitements rapides.

Jusqu'ici j'ai limités mes études au processus dits 'feedforward', parce qu'ils sont supposés être le principal corrélat neuronal de la reconnaissance ultra-rapide. Cependant, il est clair que le feedback et les effets 'top-down'

jouent un rôle clef dans le mode 'normal' de la vision. Mieux comprendre leurs rôles est mon objectif principal de post-doctorat.

6.2 On selectivity, invariance and speed in the visual system

In the introduction we explained that the neuronal responses in the visual system have three important properties. They are:

1. Selective (see Section 1.2.1, paragraph 1)
2. Invariant (see Section 1.2.1, paragraph 2)
3. Fast (see Section 1.2.2)

Although still speculative at this time, the STDP-based learning mechanisms presented in Chapters 2 and 3 could account for both the selectivity and the speed of the responses.

To be precise we have shown in Chapter 2 that, in a feedforward network mimicking the ventral pathway, a combination of a temporal coding scheme where the most strongly activated neurons fire first with Spike-Time Dependent Plasticity leads to a situation where neurons in higher order visual areas will gradually become selective to frequently occurring feature combinations. At the same time, their responses become more and more rapid. The resulting model is appealing because, like other hierarchical models Fukushima (1980); LeCun and Bengio (1998); Riesenhuber and Poggio (1999); Wallis and Rolls (1997); Rolls and Milward (2000); Stringer and Rolls (2000); Serre et al. (2007), it is able of robust object recognition without combinatorial explosion, but it can also do it fast, as has been suggested in experimental literature (Oram and Perrett, 1992; Thorpe et al., 1996; Fabre-Thorpe et al., 1998; Keysers et al., 2001; Rousselet et al., 2002; Bacon-Mace et al., 2005; Hung et al., 2005; Kirchner and Thorpe, 2006; Serre et al., 2007; Girard et al., 2007). We thus firmly believe that time-to-first spike coding and STDP are keys to understanding the remarkable efficiency of the primate visual system.

The study presented in Chapter 3 legitimizes the 'one-by-one processing' approximation of Chapter 2 by showing that even in a continuous regime, where afferents fire continuously with a constant population rate, STDP is still able to detect and learn a repeating spatio-temporal spike pattern.

The STDP models of Chapters 2 and 3 both predict that visual responses' latencies should decrease after repeated presentations of a same stimulus. In Chapter 4 I tested this prediction experimentally by inferring the visual

processing times through behavioral measures, and did find that familiarity with an image could speed up its recognition by about 25 ms. A ceiling effect was reached after a few hundred presentations. Of course this does not mean that the STDP models of Chapters 2 and 3 are true – only that they are *plausible*.

A new variant of the trace rule (Földiák, 1991) that could account for certain invariances of the responses has been proposed in Chapter 5. Although this study could be pushed further (in particular the proposed mechanism should be implemented on spiking neurons, and should be tested on higher order neurons), it shows how neurons could take advantage on the temporal continuity of the world, using a learning rule that incorporates a running average of the cell's activity over a recent past, in order to build invariant representations. The proposed rule works even in natural cluttered environments.

Taken together these results thus suggest how the visual cortex could wire itself to produce fast, selective and invariant responses. While still speculative at the time of writing the models presented here all rely on widely accepted biophysical phenomena and are thus biologically plausible.

6.3 On learning rates

It is worth mentioning that the theoretical STDP studies of Chapters 2 and 3, and the experimental study of Chapter 4 all suggest that learning takes a few hundred iterations – an order of magnitude which is compatible with the results of *in vivo* supervised learning procedures to durably change receptive field properties in cat V1 Frégnac et al. (1988); McLean and Palmer (1998); Frégnac and Shulz (1999).

6.4 On temporal coding in general

Besides these results on vision this thesis also strengthen the case for the use of temporal coding in the brain. In particular the study presented on Chapter 3 is very general: it is not restricted to vision, nor even to sensory systems. We showed that, surprisingly, the widely accepted mechanism of STDP is able to solve a very difficult computational problem: to localize a spike pattern embedded in equally dense 'distractor' spike trains. STDP thus enables some form of temporal coding, even in the absence of an explicit time reference.

This model, together with the visual system model of Chapter 2, demon-

strates how the brain could easily make use of information encoded in the spike times. Whether these spike times contain additional information with respect to the averaged firing rates is controversial. Given that the mechanisms proposed here are simple, efficient, and satisfy the known temporal constraints coming from the experimental literature, they provide a strong argument in favor of temporal coding.

This does not mean that the rates contain no information, nor that they systematically contain less information than the spike times. Rate coding is probably extensively used in the brain. However, I think temporal coding schemes do a better work at explaining what happens during transients which are probably an important aspect of neural computation, especially when rapid processing is involved.

6.5 Perspective: top-down effects and feedback

So far I have studied mainly the feedforward paths of the ventral stream, mainly because they are believed to be the principal neural correlates of 'fast recognition'. However the temporal constraints mentioned in Section 1.2.2 do not rule out all the top-down effects. For example when performing an animal/non animal ultra rapid categorization, subjects are likely to enhance the selectivity of some neurons tuned to animal features, probably located in IT. Indeed, there is experimental evidence that the selectivity to the 'relevant' features for a given recognition task can be enhanced in IT (Sigala and Logothetis, 2002) and in V4 (Bichot et al., 2005), possibly thanks to a top-down signal coming from the prefrontal cortex, thought to be involved in the categorization process. None of the models presented in this thesis take these top-down effects into account.

Furthermore, slower processes than 'fast-recognition', such as segmentation, are believed to rely extensively on feedback loops.

It is thus very clear that back-projections play a critical role in normal, every day vision, and I would like to get a better understanding of their functions during my postdoc.

6.6 On the roles of models

Sadly, models sometimes receive little interest from the experimental community. Some think that it is too early to build realistic models, that we should first gather enough experimental data. Some do not even see the point in building an artificial brain, in parallel of the real one. Here I list the main

roles of models in neuroscience, according to me.

- **To organize and synthesize the knowledge we accumulated about a given brain function.** This kind of models need not be quantitative. They aim at giving a synthetic view of one brain function, to describe what modules are involved and the way the interact. They are built from an exhaustive reviews of the corresponding experimental literature.

- **To validate a given theory by 'implementing' it (plausibility proof)**. These quantitative models strengthen the case for a given theory. For example there is a big difference between claiming that utra-rapid visual categorization can be done in a feedforward-only mode and developing a model that actually does it. The fact that the model works is a necessary condition (although not sufficient) for the theory to be true.

- **To test some hypothesis and suggest new experiments.** Sometimes models can precede experimental data. These models are speculative but, through simulations, they can make predictions and suggest new experiments to test them.

6.7 Applications

This work was partially funded by SpikeNet Technology Inc. (`http://www.spikenet-technology.com`), which is interested in artificial vision and its industrial applications. During those three years I have worked in constant interaction with the Research and Development team, particularly with Jong-Mo Allegraud and Nicolas Guilbaud. Some applications of my work are now fully operational, others are still being evaluated.

1. STDP-based visual feature learning. A successful proof-of-concept for an early version of the algorithm presented in Chapter 2 was done in 2005 with the Centre National d'Etudes Spatiales(CNES)(`http://www.cnes.fr`) on the problematic of generic object classification in SPOT 5 satellite images (ROBIN competition, `http://robin.inrialpes.fr`, Dataset #2). In 2006, the algorithm also enabled the development of a generic face detector, SNFace, based on STDP-learnt face features. Jong-Mo Allegraud has now ported my original Matlab/C code in a highly flexible C++ code that is used for commercial applications. The next step will be to learn a huge set of generic visual features from

a huge image database, which could then be used to recognize any class of objects.

2. STDP-based spike pattern learning. Applications for audiovisual processing are now being evaluated. After having converted the input into spike trains the algorithm can learn and detect repeating patterns, which may enable sound, object and motion recognition.

3. Invariance learning. This algorithm has not been used yet, but it could be useful to restrict the zones in which features are looked for in a feature-based object detection, like in Fig. 2.1

Appendix A

Papers (Peer-reviewed international journals)

A.1 Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity

Timothée Masquelier[1,2*], **Simon J. Thorpe**[1,2]

1 Centre de Recherche Cerveau et Cognition, Centre National de la Recherche Scientifique, Université Paul Sabatier, Faculté de Médecine de Rangueil, Toulouse, France, **2** SpikeNet Technology SARL, Labege, France

Spike timing dependent plasticity (STDP) is a learning rule that modifies synaptic strength as a function of the relative timing of pre- and postsynaptic spikes. When a neuron is repeatedly presented with similar inputs, STDP is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early, while postsynaptic spike latencies decrease. Here we use this learning rule in an asynchronous feedforward spiking neural network that mimics the ventral visual pathway and shows that when the network is presented with natural images, selectivity to intermediate-complexity visual features emerges. Those features, which correspond to prototypical patterns that are both salient and consistently present in the images, are highly informative and enable robust object recognition, as demonstrated on various classification tasks. Taken together, these results show that temporal codes may be a key to understanding the phenomenal processing speed achieved by the visual system and that STDP can lead to fast and selective responses.

Citation: Masquelier T, Thorpe SJ (2007) Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput Biol 3(2): e31. doi:10.1371/journal.pcbi.0030031

Introduction

Temporal constraints pose a major challenge to models of object recognition in cortex. When two images are simultaneously flashed to the left and right of fixation, human subjects can make reliable saccades to the side where there is a target animal in as little as 120–130 ms [1]. If we allow 20–30 ms for motor delays in the oculomotor system, this implies that the underlying visual processing can be done in 100 ms or less. In monkeys, recent recordings from inferotemporal cortex (IT) showed that spike counts over time bins as small as 12.5 ms (which produce essentially a binary vector with either ones or zeros) and only about 100 ms after stimulus onset contain remarkably accurate information about the nature of a visual stimulus [2]. This sort of rapid processing presumably depends on the ability of the visual system to learn to recognize familiar visual forms in an unsupervised manner. Exactly how this learning occurs constitutes a major challenge for theoretical neuroscience. Here we explored the capacity of simple feedforward network architectures that have two key features. First, when stimulated with a flashed visual stimulus, the neurons in the various layers of the system fire asynchronously, with the most strongly activated neurons firing first—a mechanism that has been shown to efficiently encode image information [3]. Second, neurons at later stages of the system implement spike timing dependent plasticity (STDP), which is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early [4,5]. We demonstrate that when such a hierarchical system is repeatedly presented with natural images, these intermediate-level neurons will naturally become selective to patterns that are reliably present in the input, while their latencies decrease, leading to both fast and informative responses. This process occurs in an entirely unsupervised way, but we then show that these intermediate features are able to support categorization.

Our network belongs to the family of feedforward hierarchical convolutional networks, as in [6–10]. To be precise, its architecture is inspired from Serre, Wolf, and Poggio's model of object recognition [6], a model that itself extends HMAX [7] and performs remarkably well with natural images. Like them, in an attempt to model the increasing complexity and invariance observed along the ventral pathway [11,12], we use a four-layer hierarchy (S1–C1–S2–C2) in which simple cells (S) gain their selectivity from a linear sum operation, while complex cells (C) gain invariance from a nonlinear max pooling operation (see Figure 1 and Methods for a complete description of our model).

Nevertheless, our network does not only rely on static nonlinearities: it uses spiking neurons and operates in the temporal domain. At each stage, the time to first spike with respect to stimulus onset (or, to be precise, the rank of the first spike in the spike train, as we will see later) is supposed to be the "key variable," that is, the variable that contains information and that is indeed read out and processed by downstream neurons. When presented with an image, the first layer's S1 cells, emulating V1 simple cells, detect edges with four preferred orientations, and the more strongly a cell is activated, the earlier it fires. This intensity–latency conversion is in accordance with recordings in V1 showing

Editor: Karl J. Friston, University College London, United Kingdom

Received November 10, 2006; **Accepted** January 2, 2007; **Published** February 16, 2007

A previous version of this article appeared as an Early Online Release on January 2, 2007 (doi:10.1371/journal.pcbi.0030031.eor).

Abbreviations: IT, inferotemporal cortex; RBF, radial basis function; ROC, receiver operator characteristic; STDP, spike timing dependent plasticity

* To whom correspondence should be addressed. E-mail: timothee.masquelier@alum.mit.edu

Author Summary

The paper describes a new biologically plausible mechanism for generating intermediate-level visual representations using an unsupervised learning scheme. These representations can then be used very effectively to perform categorization tasks using natural images. While the basic hierarchical architecture of the system is fairly similar to a number of other recent proposals, the key differences lie in the level of description that is used—individual neurons and spikes—and in the sort of coding scheme involved. Essentially, we have found that a combination of a temporal coding scheme where the most strongly activated neurons fire first with spike timing dependent plasticity leads to a situation where neurons in higher order visual areas will gradually become selective to frequently occurring feature combinations. At the same time, their responses become more and more rapid. We firmly believe that such mechanisms are a key to understanding the remarkable efficiency of the primate visual system.

that response latency decreases with the stimulus contrast [13,14] and with the proximity between the stimulus orientation and the cell's preferred orientation [15]. It has already been shown how such orientation selectivity can emerge in V1 by applying STDP on spike trains coming from retinal ON- and OFF-center cells [16], so we started our model from V1 orientation-selective cells. We also limit the number of spikes at this stage by introducing competition between S1 cells through a one-winner-take-all mechanism: at a given location—corresponding to one cortical column—only the spike corresponding to the best matching orientation is propagated (sparsity is thus 25% at this stage). Note that k-winner-take-all mechanisms are easy to implement in the temporal domain using inhibitory GABA interneurons [17].

These S1 spikes are then propagated asynchronously through the feedforward network of integrate-and-fire neurons. Note that within this time-to-first-spike framework, the maximum operation of complex cells simply consists of propagating the first spike emitted by a given group of afferents [18]. This can be done efficiently with an integrate-and-fire neuron with low threshold that has synaptic connections from all neurons in the group.

Images are processed one by one, and we limit activity to at most one spike per neuron, that is, only the initial spike wave is propagated. Before presenting a new image, every neuron's potential is reset to zero. We process various scaled versions of the input image (with the same filter size). There is one S1–C1–S2 pathway for each processing scale (not represented on Figure 1). This results in S2 cells with various receptive field sizes (see Methods). Then C2 cells take the maximum response (i.e., first spike) of S2 cells over all positions and scales, leading to position and scale invariant responses.

This paper explains how STDP can set the C1–S2 synaptic connections, leading to intermediate-complexity visual features, whose equivalent in the brain may be in V4 or IT. STDP is a learning rule that modifies the strength of a neuron's synapses as a function of the precise temporal relations between pre- and postsynaptic spikes: an excitatory synapse receiving a spike before a postsynaptic one is emitted is potentiated (long-term potentiation) whereas its strength is weakened the other way around (long-term depression) [19]. The amount of modification depends on the delay between these two events: maximal when pre- and postsynaptic spikes are close together, and the effects gradually decrease and disappear with intervals in excess of a few tens of milliseconds [20–22]. Note that STDP is in agreement with Hebb's postulate because presynaptic neurons that fired slightly before the postsynaptic neuron are those that "took part in firing it." Here we used a simplified STDP rule where the weight modification does not depend on the delay between pre- and postsynaptic spikes, and the time window is supposed to cover the whole spike wave (see Methods). We also use 0 and 1 as "soft bounds" (see Methods), ensuring the synapses remain excitatory. Several authors have studied the effect of STDP with Poisson spike trains [4,23]. Here, we demonstrate STDP's remarkable ability to detect statistical regularities in terms of earliest firing afferent patterns within visual spike trains, despite their very high dimensionality inherent to natural images.

Visual stimuli are presented sequentially, and the resulting spike waves are propagated through to the S2 layer, where STDP is used. We use restricted receptive fields (i.e., S2 cells only integrate spikes from an $s \times s$ square neighborhood in the C1 maps corresponding to one given processing scale) and weight-sharing (i.e., each *prototype* S2 cell is duplicated in retinotopic maps and at all scales). Starting with a random weight matrix (size $= 4 \times s \times s$), we present the first visual stimuli. Duplicated cells are all integrating the spike train and compete with each other. If no cell reaches its threshold, nothing happens and we process the next image. Otherwise for each prototype the first duplicate to reach its threshold is the winner. A one-winner-take-all mechanism prevents the other duplicated cells from firing. The winner thus fires and the STDP rule is triggered. Its weight matrix is updated, and the change in weights is duplicated at all positions and scales. This allows the system to learn patterns despite changes in position and size in the training examples. We also use local inhibition between different prototype cells: when a cell fires at a given position and scale, it prevents all other cells from firing later at the same scale and within an $s/2 \times s/2$ square neighborhood of the firing position. This competition, only used in the learning phase, prevents all the cells from learning the same pattern. Instead, the cell population self-organizes, each cell trying to learn a distinct pattern so as to cover the whole variability of the inputs.

If the stimuli have visual features in common (which should be the case if, for example, they contain similar objects), the STDP process will extract them. That is, for some cells we will observe convergence of the synaptic weights (by saturation), which end up being either close to 0 or to 1. During the convergence process, synapses compete for control of the timing of postsynaptic spikes [4]. The winning synapses are those through which the earliest spikes arrive (on average) [4,5], and this is true even in the presence of jitter and spontaneous activity [5] (although the model presented in this paper is fully deterministic). This "preference" for the earliest spikes is a key point since the earliest spikes, which correspond in our framework to the most salient regions of an image, have been shown to be the most informative [3]. During the learning, the postsynaptic spike latency decreases [4,5,24]. After convergence, the responses become selective (in terms of latency) [5] to visual features of intermediate complexity similar to the features used in earlier work [8]. Features can now be defined as clusters of afferents that are

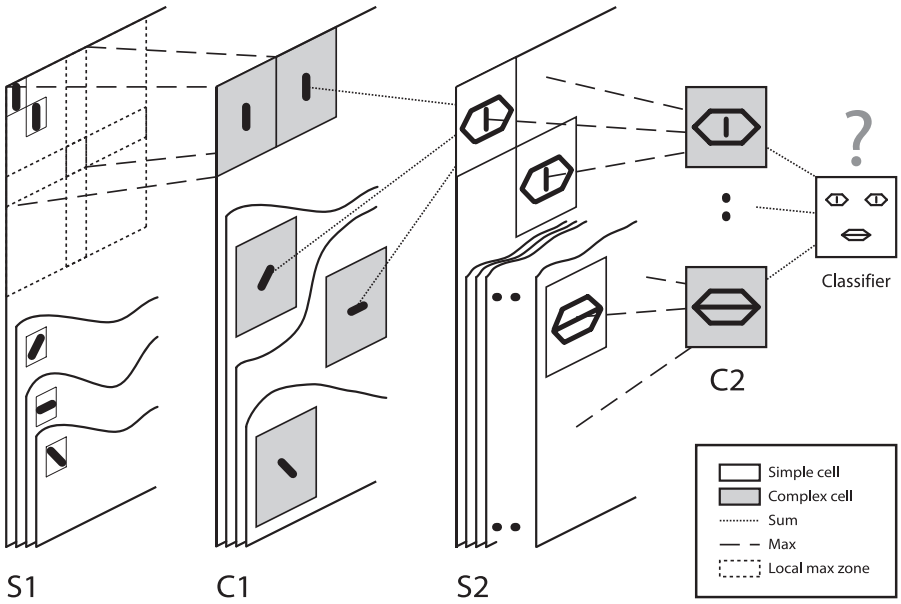


Figure 1. Overview of the Five-Layer Feedforward Spiking Neural Network

As in HMAX [7], we alternate simple cells that gain selectivity through a sum operation, and complex cells that gain shift and scale invariance through a max operation (which simply consists of propagating the first received spike). Cells are organized in retinotopic maps until the S2 layer (inclusive). S1 cells detect edges. C1 maps subsample S1 maps by taking the maximum response over a square neighborhood. S2 cells are selective to intermediate-complexity visual features, defined as a combination of oriented edges (here we symbolically represented an eye detector and a mouth detector). There is one S1–C1–S2 pathway for each processing scale (not represented). Then C2 cells take the maximum response of S2 cells over all positions and scales and are thus shift- and scale-invariant. Finally, a classification is done based on the C2 cells' responses (here we symbolically represented a face/nonface classifier). In the brain, equivalents of S1 cells may be in V1, S2 cells in V1–V2, S2 cells in V4–PIT, C2 cells in AIT, and the final classifier in PFC. This paper focuses on the learning of C1 to S2 synaptic connections through STDP.

doi:10.1371/journal.pcbi.0030031.g001

consistently among the earliest to fire. STDP detects these kinds of statistical regularities among the spike trains and creates one unit for each distinct pattern.

Results

We evaluated our STDP-based learning algorithm on two California Institute of Technology datasets, one containing faces and the other motorbikes, and a distractor set containing backgrounds, all available at http://www.vision.caltech.edu (see Figure 2 for sample pictures). Note that most of the images are not segmented. Each dataset was split into a training set, used in the learning phase, and a testing set, not seen during the learning phase but used afterward to evaluate the performance on novel images. This standard cross-validation procedure allows the measurement of the system's ability to *generalize,* as opposed to learning the specific training examples. The splits used were the same as Fergus, Perona, and Zisserman [25]. All images were rescaled to be 300 pixels in height (preserving the aspect ratio) and converted to grayscale values.

We first applied our unsupervised STDP-based algorithm on the face and motorbike training examples (separately), presented in random order, to build two sets of ten class-specific C2 features. Each C2 cell has one preferred input, defined as a combination of edges (represented by C1 cells). Note that many gray-level images may lead to this combination of edges because of the local max operation of C1 cells and because we lose the "polarity" information (i.e., which side of the edge is darker). However, we can reconstruct a representation of the set of preferred images by convolving the weight matrix with a set of kernels representing oriented bars. Since we start with random weight matrices, at the beginning of the learning process the reconstructed preferred stimuli do not make much sense. But as the cells learn, structured representations emerge, and we are usually able to identify the nature of the cells' preferred stimuli. Figures 3 and 4 show the reconstructions at various stages of learning for the face and motorbike datasets, respectively. We stopped the learning after 10,000 presentations.

Then we turned off the STDP rule and tested these STDP-obtained features' ability to support face/nonface and motorbike/nonmotorbike classification. This paper focuses more on feature extraction than on sophisticated classification methods, so we first used a very simple decision rule based on the number of C2 cells that fired with each test image, on which a threshold is applied. Such a mechanism could be easily implemented in the brain. The threshold was set at the equilibrium point (i.e., when the false positive rate equals the missed rate). In Table 1 we report good classification results with this "simple-count" scheme in terms of area under the receiver operator characteristic (ROC) and the performance rate at equilibrium point.

We also evaluated a more complicated classification scheme. C2 cells' thresholds were supposed to be infinite, and we measured the final potentials they reached after having integrated the whole spike train generated by the image. This final potential can be seen as the number of early spikes in common between a current input and a stored prototype (this contrasts with HMAX and extensions [6,7,26], where a Euclidian distance or a normalized dot product is used to measure the difference between a stored prototype

Figure 2. Sample Pictures from the Caltech Datasets

The top row shows examples of faces (all unsegmented), the middle row shows examples of motorbikes (some are segmented, others are not), and the bottom row shows examples of distractors.

doi:10.1371/journal.pcbi.0030031.g002

and a current input). Note that this potential is contrast invariant: a change in contrast will shift all the latencies but will preserve the spike order. The final potentials reached with the training examples were used to train a radial basis function (RBF) classifier (see Methods). We chose this classifier because linear combination of Gaussian-tuned units is hypothesized to be a key mechanism for generalization in the visual system [27]. We then evaluated the RBF on the testing sets. As can be seen in Table 1, performance with this "potential + RBF" scheme was better.

Using only ten STDP-learnt features, we reached on those two classes a performance that is comparable to that of Serre, Wolf, and Poggio's model, which itself is close to the best state-of-the-art computer vision systems [6]. However, their system is more generic. Classes with more intraclass variability (for example, animals) appear to pose a problem with our approach because a lot of training examples (say a few tens) of a given feature type are needed for the STDP process to learn it properly.

Our approach leads to the extraction of a small set (here ten) of highly informative class-specific features. This is in contrast with Serre et al.'s approach where many more (usually about a thousand) features are used. Their sets are more generic and are suitable for many different classes [6]. They rely on the final classifier to "select" diagnostic features and appropriately weight them for a given classification task. Here, STDP will naturally focus on what is common to the positive training set, that is, target object features. The background is generally not learned (at least not in priority), since backgrounds are almost always too different from one image to another for the STDP process to converge. Thus, we directly extract diagnostic features, and we can obtain reasonably good classification results using only a threshold on the number of detected features. Furthermore, as STDP performs vector quantization from multiple examples as opposed to "one-shot learning," it will not learn the noise, nor anything too specific to a given example, with the result that it will tend to learn archetypical features.

Another key point is the natural trend of the algorithm to learn salient regions, simply because they correspond to the earliest spikes, with the result that neurons whose receptive fields cover salient regions are likely to reach their threshold (and trigger the STDP rule) before neurons "looking" at other regions. This contrasts with more classical competitive

Table 1. Classification Results

Model	STDP Features (Simple Count)		STDP Features (Potential + RBF)		Hebbian Features		Serre, Wolf, and Poggio	
	Equilibrium Point	ROC	Equilibrium Point	ROC	Equilibrium Point	ROC	Equilibrium Point	ROC
Faces	96.5	99.1	99.1	100.0	96.9	99.7	98.2	99.8
Motorbikes	95.4	98.4	97.8	99.7	96.5	99.3	98	99.8

doi:10.1371/journal.pcbi.0030031.t001

Figure 3. Evolution of Reconstructions for Face Features

At the top is the number of postsynaptic spikes emitted. Starting from random preferred stimuli, cells detect statistical regularities among the input visual spike trains after a few hundred discharges and progressively develop selectivity to those patterns. A few hundred more discharges are needed to reach a stable state. Furthermore, the population of cells self-organizes, with each cell effectively trying to learn a distinct pattern so as to cover the whole variability of the inputs.

doi:10.1371/journal.pcbi.0030031.g003

learning approaches, where input normalization helps different input patterns to be equally effective in the learning process [28]. Note that "salient" means within our network "with well-defined contrasted edges," but saliency is a more generic concept of local differences, for example, in intensity, color, or orientations as in the model of Itti, Koch, and Niebur [29]. We could use other types of S1 cells to detect other types of saliency, and, provided we apply the same intensity–latency conversion, STDP would still focus on the most salient regions. Saliency is known to drive attention (see [30] for a review). Our model predicts that it also drives the learning. Future experimental work will test this prediction.

Of course, in real life we are unlikely to see many examples of a given category in a row. That is why we performed a second simulation, where 20 C2 cells were presented with the face, motorbike, and background training pictures in random order, and the STDP rule was applied. Figure 5 shows all the reconstructions for this mixed simulation after 20,000 presentations. We see that the 20 cells self-organized, some of them having developed selectivity to face features, and others to motorbike features. Interestingly, during the

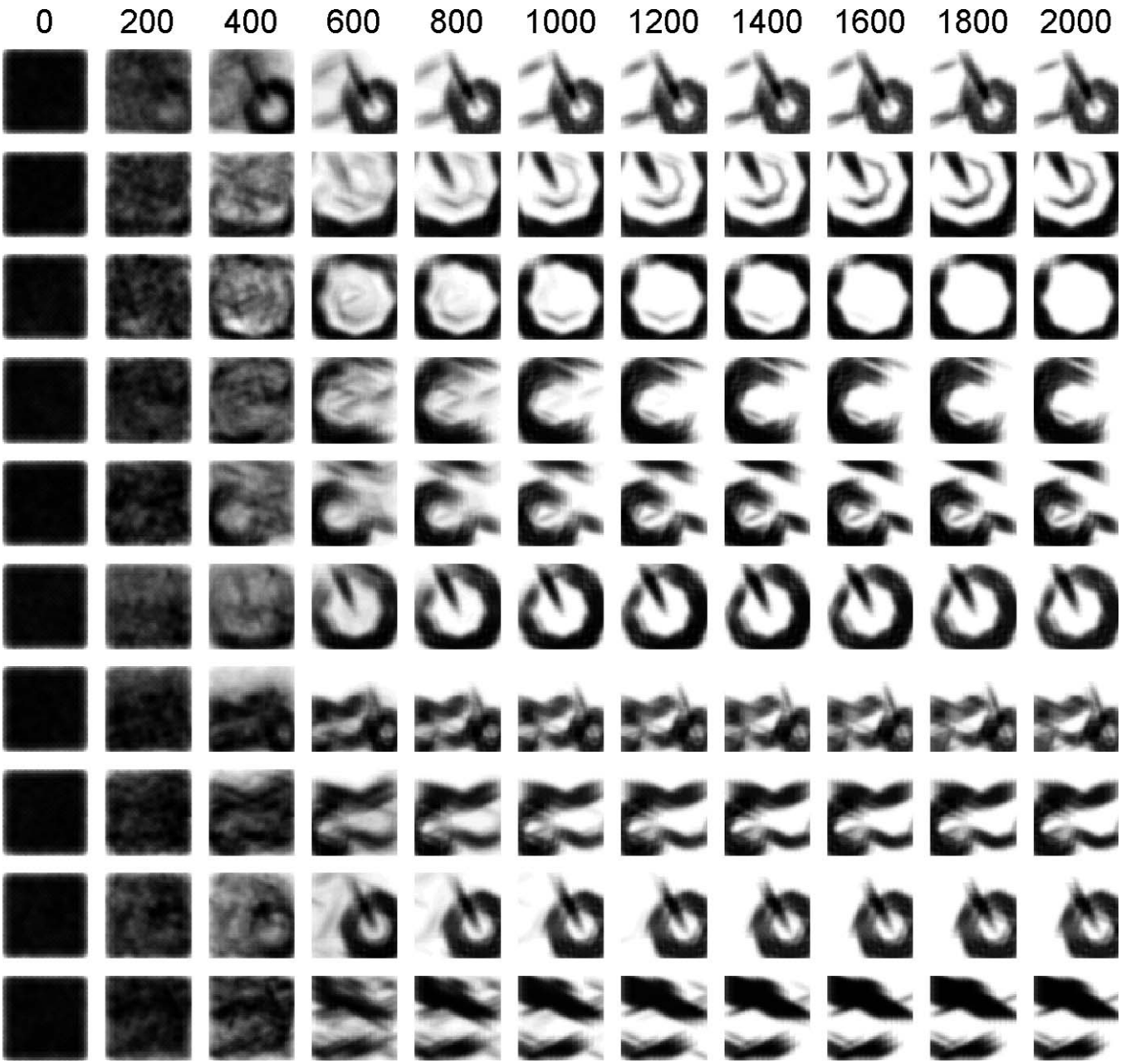


Figure 4. Evolution of Reconstructions for Motorbike Features
doi:10.1371/journal.pcbi.0030031.g004

learning process the cells rapidly showed a preference for one category. After a certain degree of selectivity had been reached, the face-feature learning was not influenced by the presentation of motorbikes (and vice versa), simply because face cells will not fire (and trigger the STDP rule) on motorbikes. Again we tested the quality of these features with a (multiclass) classification task, using an RBF network and a "one-versus-all" approach (see Methods). As before, we tested two implementations: one based on "binary detections + RBF" and one based on "potential + RBF". Note that a simple detection count cannot work here, as we need at least some supervised learning to know which feature (or feature combination) is diagnostic (or antidiagnostic) of which class. Table 2 shows the confusion matrices obtained on the testing sets for both implementations, leading, respectively, to 95.0% and 97.7% of correct classifications on average. It is worth mentioning that the "potential + RBF" system perfectly discriminated between faces and motorbikes—although both were presented in the unsupervised STDP-based learning phase.

A third type of simulation was run to illustrate the STDP learning process. For these simulations, only three C2 cells and four processing scales (71%, 50%, 35%, and 25%) were used. We let at most one cell fire at each processing scale. The rest of the parameters were strictly identical to the other simulations (see Methods). Videos S1–S3 illustrate the STDP

Table 2. Confusion Matrices

Predicted with:	STDP Features (Binary Detections)			STDP Features (Potential)			Hebbian Features		
	Face	Motorbike	Background	Face	Motorbike	Background	Face	Motorbike	Background
Actual Face	97.2	0.5	2.3	98.2	0	1.8	97.7	0	2.3
Actual Motorbike	0	95.3	4.8	0	97.5	2.5	0.3	96.3	3.5
Actual Background	3.1	4.4	92.4	0.4	2.2	97.3	4.9	3.6	91.6

doi:10.1371/journal.pcbi.0030031.t002

learning process with, respectively, faces, motorbikes, and a mix of faces, motorbikes, and background pictures. It can be seen that after convergence the STDP feature showed a good tradeoff between selectivity (very few false alarms) and invariance (most of the targets were recognized).

An interesting control is to compare the STDP learning rule with a more standard hebbian rule in this precise framework. For this purpose, we converted the spike trains coming from C1 cells into a vector of (real-valued) C1 activities X_{C1}, supposed to correspond to firing rates (see Methods). Each S2 cell was no longer modeled at the integrate-and-fire level but was supposed to respond with a (static) firing rate Y_{S2} given by the normalized dot product:

$$Y_{S2} = \frac{W_{S2} \cdot X_{C1}}{|X_{C1}|_2} \quad (1)$$

where W_{S2} is the synaptic weight vector of the S2 cell (see Methods).

The S2 cells still competed with each other, but the *k*-winner-take-all mechanisms now selected the cells with the highest firing rates (instead of the first one to fire). Only the cells whose firing rates reached a certain threshold were considered in the competition (see Methods). The winners now triggered the following modified hebbian rule (instead of STDP):

$$\delta W_{S2} = a \cdot Y_{S2} \cdot (X_{C1} - W_{S2}), \quad (2)$$

where a decay term has been added to keep the weight vector bounded (however, the rule is still local, unlike an explicit weight normalization). Note that this precaution was not needed in the STDP case because competition between synapse naturally bounds the weight vector [4]. The rest of the network is strictly identical to the STDP case.

Figure 6 shows the reconstruction of the preferred stimuli for the ten C2 cells after 10,000 presentations for the face stimuli (Figure 6, top) and the motorbikes stimuli (Figure 6, top). Again we can usually recognize the face and motorbike parts to which the cells became selective (even though the reconstructions look fuzzier than in the STDP case because the final weights are more graded). We also tested the ability of these hebbian-obtained features to support face/nonface and motorbike/nonmotorbike classification once fed into an RBF, and the results are shown in Table 1 (last column). We also evaluated the hebbian features with the multiclass setup. Twenty cells were presented with the same mix of face, motorbike, and background pictures as before. Figure 7 shows the final reconstructions after 20,000 presentations, and Table 2 shows the confusion matrix (last columns).

The main conclusion is that the modified hebbian rule is also able to extract pertinent features for classification (although performance on these tests appears to be slightly

Figure 5. Final Reconstructions for the 20 Features in the Mixed Case

The 20 cells self-organized, some having developed selectivity to face features, and some to motorbike features.

doi:10.1371/journal.pcbi.0030031.g005

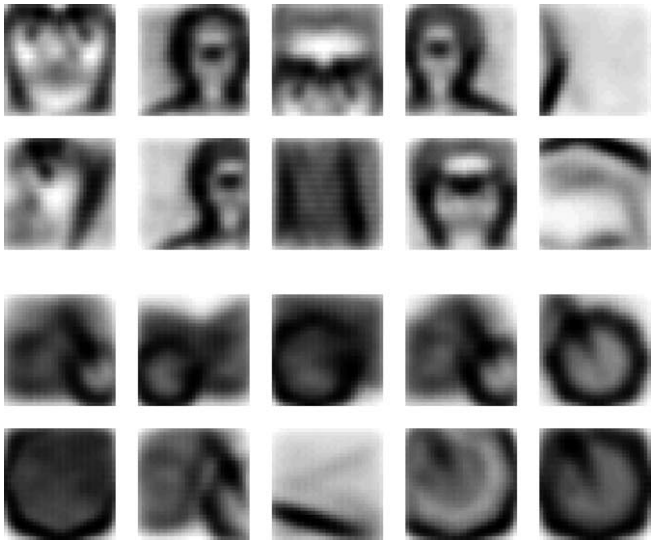


Figure 6. Hebbian Learning

(Top) Final reconstructions for the ten face features.
(Bottom) The ten motorbike features.

doi:10.1371/journal.pcbi.0030031.g006

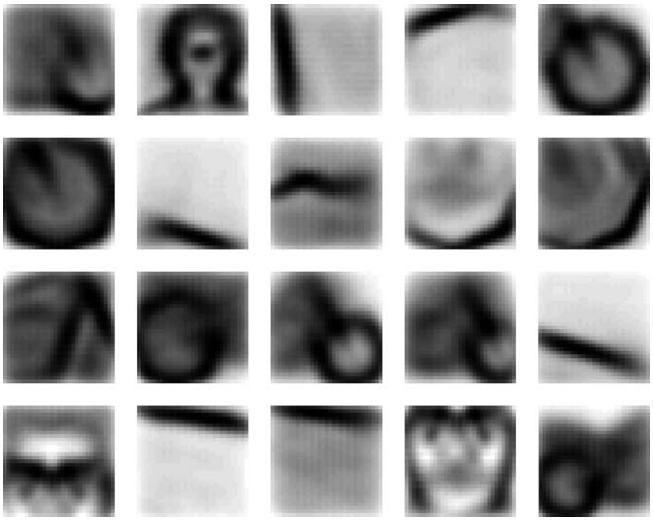


Figure 7. Hebbian Learning: Final Reconstructions for the 20 Features in the Mixed Case

As with STDP-based learning, the 20 cells self-organized, some having developed selectivity to face features, and some to motorbike features.
doi:10.1371/journal.pcbi.0030031.g007

worse). This is not very surprising as STDP can be seen as a hebbian rule transposed in the temporal domain, but it was worth checking. Where STDP would detect (and create selectivity to) sets of units that are consistently among the first one to fire, the hebbian rule detects (and creates selectivity to) sets of units that consistently have the highest firing rates. However, we believe the temporal framework is a better description of what really happens at the neuronal level, at least in ultrarapid categorization tasks. Furthermore, STDP also explains how the system becomes faster and faster with training, since the neurons learn to decode the first information available at their afferents' level (see also Discussion).

Discussion

While the ability of hierarchical feedforward networks to support classification is now reasonably well established (e.g., [6–8,10]), how intermediate-complexity features can be learned remains an open problem, especially with cluttered images. In the original HMAX model, S2 features were not learned but were manually hardwired [7]. Later versions used huge sets of random crops (say 1,000) taken from natural images and used these crops to "imprint" S2 cells [6]. This approach works well but is costly since redundancy is very high between features, and many features are irrelevant for most (if not all) of the tasks. To select only pertinent features for a given task, Ullman proposed an interesting criterion based on mutual information [8], leaving the question of possible neural implementation open. LeCun showed how visual features in a convolutional network could be learned in a supervised manner using back-propagation [10], without claiming this algorithm was biologically plausible. Although we may occasionally use supervised learning to create a set of features suitable for a particular recognition task, it seems unrealistic that we need to do that each time we learn a new class. Here we took another approach: one layer with unsupervised competitive learning is used as input for a second layer with supervised learning. Note that this kind of hybrid scheme has been found to learn much faster than a two-layer backpropagation network [28].

Our approach is a bottom-up one: instead of intuiting good image-processing schemes and discussing their eventual neural correlates, we took known biological phenomena that occur at the neuronal level, namely integrate-and-fire and STDP, and observed where they could lead at a more integrated level. The role of the simulations with natural images is thus to provide a "plausibility proof" that such mechanisms could be implemented in the brain.

However, we have made four main simplifications. The first one was to propagate input stimuli one by one. This may correspond to what happens when an image is flashed in an ultrarapid categorization paradigm [1], but normal visual perception is an ongoing process. However, every 200 ms or 300 ms we typically perform a saccade. The processing of each of these discrete "chunks" seems to be optimized for rapid execution [31], and we suggest that much can be done with the feedforward propagation of a single spike wave. Furthermore, even when fixating, our eyes are continuously making microsaccades that could again result in repetitive waves of activation. This idea is in accordance with electrophysiological recordings showing that V1 neuron activity is correlated with microsaccades [32]. Here we assumed the successive waves did not interfere, which does not seem too unreasonable given that the neuronal time constants (integration, leak, STDP window) are in the range of a few tens of milliseconds whereas the interval between saccades and microsaccades is substantially longer. It is also possible that extraretinal signals suppress interference by shutting down any remaining activity before propagating the next wave. Note that this simplification allows us to use nonleaky integrate-and-fire neurons and an infinite STDP time window. More generally, as proposed by Hopfield [33], waves could be generated by population oscillations that would fire one cell at a time in advance of the maximum of the oscillation, which increases with the inputs the cell received. This idea is in accordance with recordings in area 17 of cat visual cortex showing that suboptimal cells reveal a systematic phase lag relative to optimally stimulated cells [34].

The second simplification we have made is to use only five layers (including the classification layer), whereas processing in the ventral stream involves many more layers (probably about ten), and complexity increases more slowly than suggested here. However, STDP as a way to combine simple features into more complex representations, based on statistical regularities among earliest spike patterns, seems to be a very efficient learning rule and could be involved at all stages.

The third main simplification we have made consists of using restricted receptive fields and weight sharing, as do most of the bio-inspired hierarchical networks [6–10] (networks using these techniques are called *convolutional networks*). We built shift and scale invariance by structure (and not by training) by duplicating S1, C1, and S2 cells at all positions and scales. This is a way to reduce the number of free parameters (and therefore the VC dimension [35]) of the network by incorporating prior information into the network design: responses should be scale- and shift-invariant. This greatly reduces the number of training

examples needed. Note that this technique of weight sharing could be applied to other transformations than shifting and scaling, for instance, rotation and symmetry. However, it is difficult to believe that the brain could really use weight sharing since, as noted by Földiák [36], updating the weights of all the simple units connected to the same complex unit is a nonlocal operation. Instead, he suggested that at least the low-level features could be learned locally and independently. Subsequently, cells with similar preferred stimulus may connect adaptively to the same complex cell, possibly by detecting correlation across time thanks to a trace rule [36]. Wallis, Rolls, and Milward successfully implemented this sort of mechanism in a multilayered hierarchical network called Vis-Net [37,38]; however, performance after learning objects from unsegmented natural images was poor [39]. Future work will evaluate the use of local learning and adaptative complex pooling in our network, instead of exact weight sharing. Learning will be much slower but should lead to similar STDP features. Note that it seems that monkeys can recognize high-level objects at scales and positions that have not been experienced previously [2,40]. It could be that in the brain local learning and adaptative complex pooling are used up to a certain level of complexity, but not for high-level objects. These high-level objects could be represented with a combination of simpler features that would already be shift- and scale-invariant. As a result, there would be less need for spatially specific representations for high-level objects.

The last main simplification we have made is to ignore both feedback loops and top-down influences. While normal, everyday vision extensively uses feedback loops, the temporal constraints almost certainly rule them out in an ultrarapid categorization task [41]. The same cannot be said about the top-down signals, which do not depend directly on inputs. For example, there is experimental evidence that the selectivity to the "relevant" features for a given recognition task can be enhanced in IT [42] and in V4 [43], possibly thanks to a top-down signal coming from the prefrontal cortex, thought to be involved in the categorization process. These effects, for example, modeled by Szabo et al. [44], are not taken into account here.

Despite these four simplifications, we think our model captures two key mechanisms used by the visual system for rapid object recognition. The first one is the importance of the first spikes for rapidly encoding the most important information about a visual stimulus. Given the number of stages involved in high-level recognition and the short latencies of selective responses recorded in monkeys' IT [2], the time window available for each neuron to perform its computation is probably about 10–20 ms [45] and will rarely contain more than one or two spikes. The only thing that matters for a neuron is whether an afferent fires early enough so that the presynaptic spike falls in the critical time window, while later spikes cannot be used for ultrarapid categorization. At this point (but only at this point), we have to consider two hypotheses: either presynaptic spike times are completely stochastic (for example, drawn from a Poisson distribution), or they are somewhat reliable. The first hypothesis causes problems since the first presynaptic spikes (again the only ones taken into account) will correspond to a subset of the afferents that is essentially random, and will not contain much information about their real activities [46]. A solution to this problem is to use populations of redundant neurons (with similar selectivity) to ensure the first presynaptic spikes do correspond on average to the most active populations of afferents. In this work we took the second hypothesis, assuming the time to first spike of the afferents (or, to be precise, their firing order) was reliable and did reflect a level of activity. This second hypothesis receives experimental support. For example, recent recordings in monkeys show that IT neurons' responses in terms of spike count *close to stimulus onset* (100–150 ms time bin) seem to be too reliable to be fit by a typical Poisson firing rate model [47]. Another recent electrophysiological study in monkeys showed that IT cell's latencies do contain information about the nature of a visual stimulus [48]. There is also experimental evidence for precise spike time responses in V1 and in many other neuronal systems (see [49] for a review).

Very interestingly, STDP provides an efficient way to develop selectivity to first spike patterns, as shown in this work. After convergence, the potential reached by an STDP neuron is linked to the number of early spikes in common between the current input and a stored prototype. This "early spike" versus "later spike" neural code (while the spike order within each bin does not matter) has not only been proven robust enough to perform object recognition in natural images but is fast to read out: an accurate response can be produced when only the earliest afferents have fired. The use of such a mechanism at each stage of the ventral stream could account for the phenomenal processing speed achieved by the visual system.

Materials and Methods

Here is a detailed description of the network, the STDP model, and the classification methods.

S1 cells. S1 cells detect edges by performing a convolution on the input images. We are using 5×5 convolution kernels, which roughly correspond to Gabor filters with wavelength of 5 (i.e., the kernel contains one period), effective width 2, and four preferred orientations: $\pi/8$, $\pi/4 + \pi/8$, $\pi/2 + \pi/8$, and $3\pi/4 + \pi/8$ ($\pi/8$ is there to avoid focusing on horizontal and vertical edges, which are seldom diagnostic). We apply those filters to five scaled versions of the original image: 100%, 71%, 50%, 35%, and 25%. There are thus $4 \times 5 = 20$ S1 maps. S1 cells emit spikes with a latency that is inversely proportional to the absolute value of the convolution (the response is thus invariant to an image negative operation). We also limit activity at this stage: at a given processing scale and location, only the spike corresponding to the best matching orientation is propagated.

C1 cells. C1 cells propagate the first spike emitted by S1 cells in a 7×7 square of a given S1 map (which corresponds to one preferred orientation and one processing scale). Two adjacent C1 cells in a C1 map correspond to two 7×7 squares of S1 cells shifted by six S1 cells (and thus overlap of one S1 row). C1 maps thus subsample S1 maps. To be precise, neglecting the side effects, there are $6 \times 6 = 36$ times fewer C1 cells than S1 cells. As proposed by Riesenhuber and Poggio [7], this maximum operation is a biologically plausible way to gain local shift invariance. From an image processing point of view, it is a way to perform subsampling within retinotopic maps without flattening high spatial frequency peaks (as would be the case with local averaging).

We also use a local lateral inhibition mechanism at this stage: when a C1 cell emits a spike, it increases the latency of its neighbors within an 11×11 square in the map with the same preferred orientation and the same scale. The percentage of latency increase decreases linearly with the distance from the spike, from 15% to 5%. As a result, if a region is clearly dominated by one orientation, cells will inhibit each other and the spike train will be globally late and thus unlikely to be "selected" by STDP.

S2 cells. S2 cells correspond to intermediate-complexity visual features. Here we used ten prototype S2 cell types, and 20 in the mixed simulation. Each prototype cell is duplicated in five maps (weight sharing), each map corresponding to one processing scale. Within those maps, the S2 cells can integrate spikes only from the four C1 maps of the corresponding processing scale. The receptive field size is 16×16 C1 cells (neglecting the side effects; this leads to 96 $\times$ 96 S1 cells, and the corresponding receptive field size in the original image is [96 / processing scale]^2). C1–S2 synaptic connections are set by STDP.

Note that we did not use a leakage term. In the brain, by progressively resetting membrane potentials toward their resting levels, leakiness will decrease the interference between two successive spike waves. In our model we process spike waves one by one and reset all the potentials before each propagation, and so leaks are not needed.

Finally, activity is limited at this stage: a *k*-winner-take-all strategy ensures at most two cells that can fire for each processing scale. This mechanism, only used in the learning phase, helps the cells to learn patterns with different real sizes. Without it, there is a natural bias toward "small" patterns (i.e., large scales), simply because corresponding maps are larger, and so likeliness of firing with random weights at the beginning of the STDP process is higher.

C2 cells. Those cells take for each prototype the maximum response (i.e., first spike) of corresponding S2 cells over all positions and processing scales, leading to ten shift- and scale-invariant cells (20 in the mixed case).

STDP model. We used a simplified STDP rule:

$$\begin{cases} \Delta w_{ij} = a^+ . w_{ij} . (1 - w_{ij}) & if \quad t_j - t_i \leq 0 \\ \Delta w_{ij} = a^- . w_{ij} . (1 - w_{ij}) & if \quad t_j - t_i > 0 \end{cases} \tag{3}$$

where i and j refer, respectively, to the post- and presynaptic neurons, t_i and t_j are the corresponding spike times, Δw_{ij} is the synaptic weight modification, and a^+ and a^- are two parameters specifying the amount of change. Note that the weight change does not depend on the exact $t_i - t_j$ value, but only on its sign. We also used an infinite time window. These simplifications are equivalent to assuming that the intensity–latency conversion of S1 cells compresses the whole spike wave in a relatively short time interval (say, 20–30 ms), so that all presynaptic spikes necessarily fall close to the postsynaptic spike time, and the change decrease becomes negligible. In the brain, this change decrease and the limited time window are crucial: they prevent different spike waves coming from different stimuli from interfering in the learning process. In our model, we propagate stimuli one by one, so these mechanisms are not needed. Note that with this simplified STDP rule only the *order* of the spikes matters, not their precise timings. As a result, the intensity–latency conversion function of S1 cells has no impact, and any monotonously decreasing function gives the same results.

The multiplicative term $w_{ij} \cdot (1 - w_{ij})$ ensures the weight remains in the range [0,1] (excitatory synapses) and implements a soft bound effect: when the weight approaches a bound, weight changes tend toward zero.

We also applied long-term depression to synapses through which no presynaptic spike arrived, exactly as if a presynaptic spike had arrived after the postsynaptic one. This is useful to eliminate the noise due to original random weights on synapses through which presynaptic spikes never arrive.

As the STDP learning progresses, we increase a^+ and $|a^-|$ To be precise, we start with $a^+ = 2^{-6}$ and multiply the value by 2 every 400 postsynaptic spikes, until a maximum value of 2^{-2}. a^- is adjusted so as to keep a fixed a^+/a^- ratio (–4/3). This allows us to accelerate convergence when the preferred stimulus is somewhat "locked," whereas directly using high learning rates with the random initial weights leads to erratic results.

We used a threshold of 64 (= 1/4 $\times$ 16 $\times$ 16). Initial weights are randomly generated, with mean 0.8 and standard deviation 0.05.

Classification setup. We used an RBF network. In the brain, this classification step may be done in the PFC using the outputs of IT. Let X be the vector of C2 responses (containing either binary detections with the first implementation or final potentials with the second one). This kind of classifier computes an expression of the form:

$$f(X) = \sum_{i=1}^{N} c_i \cdot e^{-\frac{(X - X_i)^2}{2\sigma^2}} \tag{4}$$

and then classifies based on whether or not $f(X)$ reaches a threshold. Supervised learning at this stage involves adjusting the synaptic weights c so as to minimize a (regularized) error on the training set [27]. The X_i correspond to C2 responses for some training examples (1/4 of the training set randomly selected). The full training set was used to learn the c_i. We used $\sigma = 2$ and $\lambda = 10^{-12}$ (regularization parameter).

The multiclass case was handled with a "one-versus-all approach." If n is the number of classes (here, three), n RBF classifiers of the kind "class I" versus "all other classes" are trained. At the time of testing, each one of the n classifiers emits a (real-valued) prediction that is linked to the probability of the image belonging to its category. The assigned category is the one that corresponds to the highest prediction value.

Hebbian learning. The spike trains coming from C1 cells were converted into real-valued activities (supposed to correspond to firing rates) by taking the inverse of the first spikes' latencies (note that these activities do not correspond exactly to the convolution values because of the local lateral inhibition mechanism of layer C1). The activities (or firing rates) of S2 units were computed as:

$$Y_{S2} = \frac{W_{S2} \cdot X_{C1}}{|X_{C1}|_2} \tag{5}$$

where W_{S2} is the synaptic weight vector of the S2 cell. Note that the normalization causes an S2 cell to respond maximally when the input vector X_{C1} is collinear to its weight vector W_{S2} (neural circuits for such normalization have been proposed in [27]). Hence W_{S2} (or any vector collinear to it) is the preferred stimulus of the S2 cell. With another stimulus X_{C1} the response is proportional to the cosine between W_{S2} and X_{C1}. This kind of tuning has been used in extensions of HMAX [26]. It is similar to the Gaussian tuning of the original HMAX [7], but it is invariant to the norm of the input (i.e., multiplying the input activities by 2 has no effect on the response), which allows us to remain contrast-invariant (see also [26] for a comparison between the two kinds of tuning).

Only the cells whose activities were above a threshold were considered in the competition process. It was found useful to use individual adaptative thresholds: each time a cell was among the winners, its threshold was set to 0.91 times its activity (this value was tuned to get approximately the same number of weight updates as with STDP). The competition mechanism was exactly the same as before, except that it selected the most active units and not the first one to fire. The winners' weight vectors were updated with the following modified hebbian rule:

$$\delta W_{S2} = a \cdot Y_{S2} \cdot (X_{C1} - W_{S2}) \tag{6}$$

a is the learning rate. It was found useful to start with a small learning rate (0.002) and to geometrically increase it every ten iterations. The geometric ratio was set to reach a learning rate of 0.02 after 2,000 iterations, after which the learning rate stayed constant.

Differences from the model of Serre, Wolf, and Poggio. Here we summarize the differences between our model and their model [6] in terms of architecture (leaving the questions of learning and temporal code aside).

We process various scaled versions of the input image (with the same filter size), instead of using various filter sizes on the original image: S1 level, only the best matching orientation is propagated; C1 level, we use lateral inhibition (see above); S2 level, the similarity between a current input and the stored prototype is linked to the number of early spikes in common between the corresponding spike trains, while Serre et al. use the Euclidian distance between the corresponding patches of C1 activities.

We used an RBF network and not a Support Vector Machine.

Supporting Information

Video S1. Face-Feature Learning

Here we presented the face-training examples in random order, propagated the corresponding spike waves, and applied the STDP rule. At the top of the screen, the input image is shown, with red, green, or blue squares indicating the receptive fields of the cells that fired (if any). At the bottom of the screen, we reconstructed the preferred stimuli of the three C2 cells. Above each reconstruction, the number of postsynaptic spikes emitted is shown with the corresponding color. The red, green, and blue cells develop selectivity to a view of, respectively, the bust, the head, and the face.

Found at doi:10.1371/journal.pcbi.0030031.sv001 (3.3 MB MOV).

Video S2. Motorbike-Feature Learning

The red cell becomes selective to the front part of a motorbike, while the green and blue cells both become selective to the wheels.

Found at doi:10.1371/journal.pcbi.0030031.sv002 (6.8 MB MOV).

Video S3. Mixed Case

The training set consisted of 200 face pictures, 200 motorbike pictures, and 200 background pictures. Notice that the red cell becomes selective to faces and the blue cell to heads, while the green cell illustrates how a given feature (round shape) can be shared by two categories.

Found at doi:10.1371/journal.pcbi.0030031.sv003 (7.6 MB MOV).

Acknowledgments

We thank Thomas Serre and Rufin VanRullen for reading the manuscript and making comments.

Author contributions. TM and SJT conceived and designed the experiments, TM performed the experiments and analyzed the data, and TM and SJT wrote the paper.

Funding. This research was supported by CNRS, STREP Decisions-in-Motion (IST-027198), and SpikeNet Technology SARL.

Competing interests. The authors have declared that no competing interests exist.

References

1. Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. Vision Res 46: 1762–1776.
2. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310: 863–866.
3. VanRullen R, Thorpe SJ (2001) Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. Neural Comput 13: 1255–1283.
4. Song S, Miller KD, Abbott LF (2000) Competitive hebbian learning through spike-timing–dependent synaptic plasticity. Nat Neurosci 3: 919–926.
5. Guyonneau R, VanRullen R, Thorpe SJ (2005) Neurons tune to the earliest spikes through STDP. Neural Comput 17: 859–879.
6. Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. CVPR 2: 994–1000.
7. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2: 1019–1025.
8. Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5: 682–687.
9. Fukushima K (1980) Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36: 193–202.
10. LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. In: Arbib MA, editor. The handbook of brain theory and neural networks. Cambridge (Massachusetts): MIT Press. pp. 255–258.
11. Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71: 856–867.
12. Oram MW, Perrett DI (1994) Modeling visual recognition from neurobiological constraints. Neural Networks 7: 945–972.
13. Albrecht DG, Geisler WS, Frazor RA, Crane AM (2002) Visual cortex neurons of monkeys and cats: Temporal dynamics of the contrast response function. J Neurophysiol 88: 888–913.
14. Gawne TJ, Kjaer TW, Richmond BJ (1996) Latency: Another potential code for feature binding in striate cortex. J Neurophysiol 76: 1356–1360.
15. Celebrini S, Thorpe S, Trotter Y, Imbert M (1993) Dynamics of orientation coding in area V 1 of the awake primate. Vis Neurosci 10: 811–825.
16. Delorme A, Perrinet L, Samuelides M, Thorpe SJ (2000) Networks of Integrate-and-Fire Neurons using Rank Order Coding B: Spike Timing Dependent Plasticity and Emergence of Orientation Selectivity. Neurocomputing 38–40: 539–545.
17. Thorpe SJ (1990) Spike arrival times: A highly efficient coding scheme for neural networks. In: Eckmiller R, Hartmann G, Hauske G, editors. Parallel processing in neural systems and computers. Amsterdam: Elsevier. pp. 91–94.
18. Rousselet GA, Thorpe SJ, Fabre-Thorpe M (2003) Taking the MAX from neuronal responses. Trends Cogn Sci 7: 99–102.
19. Markram H, Lubke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275: 213–215.
20. Bi GQ, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18: 10464–10472.
21. Zhang LI, Tao HW, Holt CE, Harris WA, Poo M (1998) A critical window for cooperation and competition among developing retinotectal synapses. Nature 395: 37–44.
22. Feldman DE (2000) Timing-based LTP and LTD at vertical inputs to layer II /III pyramidal cells in rat barrel cortex. Neuron 27: 45–56.
23. VanRossum MCW, Bi GQ, Turrigiano GG (2000) Stable Hebbian learning from spike timing-dependent plasticity. J Neurosci 20: 8812–8821.
24. Gerstner W, Kistler WM (2002) Learning to be fast: Spiking neuron models. Cambridge University Press. pp. 421–432.
25. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. CVPR 2: 264–271.
26. Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, et al. (2005) A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Cambridge (Massachusetts): Massachusetts Institute of Technology CBCL Paper #259/AI Memo #2005–036.
27. Poggio T, Bizzi E (2004) Generalization in vision and motor control. Nature 431: 768–774.
28. Rolls ET, Deco G (2002) Computational neuroscience of vision. Oxford: Oxford University Press. 592 p.
29. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20: 1254–1259.
30. Treue S (2003) Abstract visual attention: The where, what, how and why of saliency. Curr Opin Neurobiol 13: 428–432.
31. Uchida N, Kepecs A, Mainen ZF (2006) Seeing at a glance, smelling in a whiff: Rapid forms of perceptual decision making. Nat Rev Neurosci 7: 485–491.
32. Martinez-Conde S, Macknik SL, Hubel DH (2000) Microsaccadic eye movements and firing of single cells in the striate cortex of macaque monkeys. Nat Neurosci 3: 251–258.
33. Hopfield JJ (1995) Pattern recognition computation using action potential timing for stimulus representation. Nature 376: 33–36.
34. König P, Engel AK, Roelfsema PR, Singer W (1995) How precise is neuronal synchronization? Neural Comput 7: 469–485.
35. Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theor Probab Appl 17: 264–280.
36. Földiák P (1991) Learning invariance from transformation sequences. Neural Comput 3: 194–200.
37. Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. Prog Neurobiol 51: 167–194.
38. Rolls ET, Milward T (2000) A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. Neural Comput 12: 2547–2572.
39. Stringer SM, Rolls ET (2000) Position invariant recognition in the visual system with cluttered environments. Neural Networks 13: 305–315.
40. Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5: 552–563.
41. Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381: 520–522.
42. Sigala N, Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415: 318–320.
43. Bichot NP, Rossi AF, Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area v4. Science 308: 529–534.
44. Szabo M, Stetter M, Deco G, Fusi S, Giudice PD, et al. (2006) Learning to attend: Modeling the shaping of selectivity in infero-temporal cortex in a categorization task. Biol Cybern 94: 351–365.
45. Thorpe SJ, Imbert M (1989) Biological constraints on connectionist modelling. In: Pfeifer R, Schreter Z, Fogelman-Soulié F, Steels L, editors. Connectionism in perspective. Amsterdam: Elsevier. pp. 63–92.
46. Gautrais J, Thorpe S (1998) Rate coding versus temporal order coding: A theoretical approach. Biosystems 48: 57–65.
47. Amarasingham A, Chen TL, Geman S, Harrison MT, Sheinberg DL (2006) Spike count reliability and the Poisson hypothesis. J Neurosci 26: 801–809.
48. Kiani R, Esteky H, Tanaka K (2005) Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. J Neurophysiol 94: 1587–1596.
49. VanRullen R, Guyonneau R, Thorpe SJ (2005) Spike times make sense. Trends Neurosci 28: 1–4.

A.2 Spike Timing Dependent Plasticity Finds the Start of Repeating Patterns in Continuous Spike Trains

OPEN ACCESS Freely available online

PLoS one

Spike Timing Dependent Plasticity Finds the Start of Repeating Patterns in Continuous Spike Trains

Timothée Masquelier[1,2]*, Rudy Guyonneau[1,2], Simon J. Thorpe[1,2]

1 Centre de Recherche Cerveau et Cognition, Université Toulouse 3, Centre National de la Recherche Scientifique (CNRS), Faculté de Médecine de Rangueil, Toulouse, France, 2 SpikeNet Technology SARL, Prologue 1 La Pyrénéenne, Labège, France

Experimental studies have observed Long Term synaptic Potentiation (LTP) when a presynaptic neuron fires shortly before a postsynaptic neuron, and Long Term Depression (LTD) when the presynaptic neuron fires shortly after, a phenomenon known as Spike Timing Dependant Plasticity (STDP). When a neuron is presented successively with discrete volleys of input spikes STDP has been shown to learn 'early spike patterns', that is to concentrate synaptic weights on afferents that consistently fire early, with the result that the postsynaptic spike latency decreases, until it reaches a minimal and stable value. Here, we show that these results still stand in a continuous regime where afferents fire continuously with a constant population rate. As such, STDP is able to solve a very difficult computational problem: to localize a repeating spatio-temporal spike pattern embedded in equally dense 'distractor' spike trains. STDP thus enables some form of temporal coding, even in the absence of an explicit time reference. Given that the mechanism exposed here is simple and cheap it is hard to believe that the brain did not evolve to use it.

Citation: Masquelier T, Guyonneau R, Thorpe SJ (2008) Spike Timing Dependent Plasticity Finds the Start of Repeating Patterns in Continuous Spike Trains. PLoS ONE 3(1): e1377. doi:10.1371/journal.pone.0001377

INTRODUCTION

Electrophysiologists report the existence of repeating spatio-temporal spike patterns with millisecond precision, both in vitro and in vivo, lasting from a few tens of ms to several seconds[1–3]. In this study we assess the difficult problem of detecting them, and suggest how neurons could solve it. The problem is made particularly difficult when only a fraction of the recorded neurons are involved in the pattern. Fig. 1 illustrates such a situation. There is a pattern of spikes (indicated by the red dots) that repeats at irregular intervals, but is hidden within the variable background firing of the whole population (shown in blue). The problem is made hard because nothing in terms of population firing rate characterizes the periods when the pattern is present, nor is there anything unusual about the firing rates of the neurons involved in the pattern. In such a situation detecting the pattern clearly requires taking the spike times into account. However direct comparison of each spike time to one another over the entire recording period and across the entire set of afferents is extremely computationally expensive. In this article we will see how a single neuron equipped with STDP can solve the problem in a different manner, taking advantage of the fact that a pattern is a succession of spike coincidences.

STDP is now a widely accepted physiological mechanism of activity-driven synaptic regulation. It has been observed extensively in vitro[4–7], and more recently in vivo in Xenopus's visual system[8,9], in the locust's mushroom body[10], and in the rat's visual cortex[11] and barrel cortex[12]. An exponential update rule fits well the synaptic modifications observed experimentally[13] (see Fig. 2). Very recently, it has also been shown that cortical reorganization in cat primary visual cortex is in accordance with STDP[14]. Note that STDP is in agreement with Hebb's postulate because it reinforces the connections with the presynaptic neurons that fired slightly before the postsynaptic neuron, which are those that 'took part in firing it'. It thereby reinforces causality links.

When a neuron is presented successively with similar volleys of input spikes STDP is known to have the effect of concentrating synaptic weights on afferents that consistently fire early, with the result that the postsynaptic spike latency decreases[15–18]. This theoretical observation is in accordance with recordings in rat's hippocampus showing that the so called 'place cells' fire earlier – relative to the cycle of the theta oscillation in hippocampus – after the animal has repeatedly traversed the corresponding area[19]. STDP has also been studied in an oscillatory mode, and was shown to be able to select only phase-locked inputs among a broad population with random phases, turning the postsynaptic neuron into a coincidence detector[20].

The main limitation of these studies is the assumption that the input spikes arrive in discrete volleys (sometimes also called 'spike waves'). They assume an explicit time reference – usually the presentation of a stimulus[15,17,18], or the maximum (or minimum) of an oscillatory drive[20,21] – that allows the specification of a time-to-first spike (or latency) for the afferents, which could be used by the brain to encode information[22,23]. Activity between the volleys is assumed to be spontaneous and much weaker. Furthermore, many studies[15,17,20] also require the pattern to be present in all volleys for the STDP to learn it, that is no 'distractor' volleys are inserted between pattern presentations. But what happens when the population of afferents is continuously firing with a constant population firing rate, so that no explicit time reference is available? Is STDP still able to find and learn spike patterns among the inputs? Is the learning robust if, more

Academic Editor: Olaf Sporns, Indiana University, United States of America

Received October 8, 2007; **Accepted** December 7, 2007; **Published** January 2, 2008

Funding: This project was supported by the CNRS, STREP Decisions-in-Motion (IST-027198), ANR projects Natstats and Hearing in Time, and SpikeNet Technology SARL.

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: timothee.masquelier@alum.mit.edu

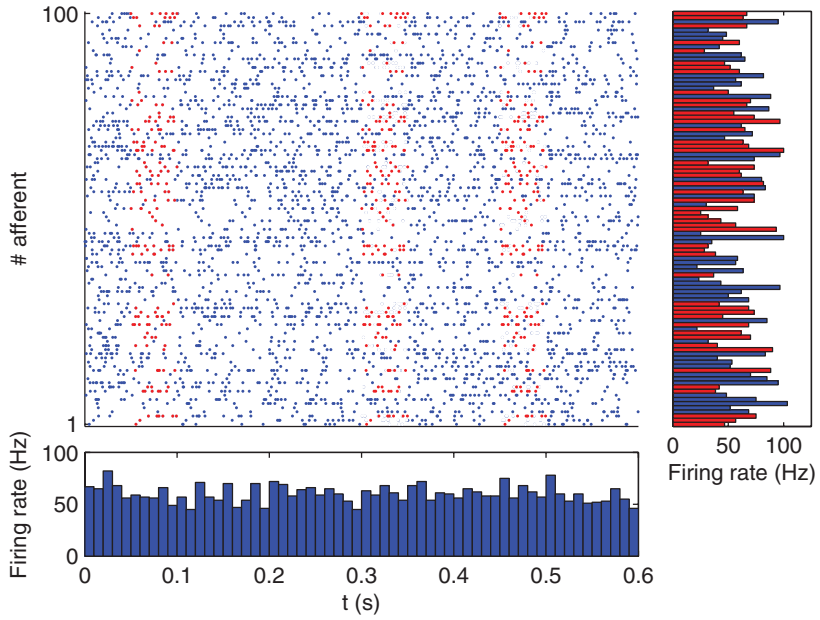


Figure 1. Spatio-temporal spike pattern. Here we show in red a repeating 50 ms long pattern that concerns 50 afferents among 100. The bottom panel plots the population-averaged firing rates over 10 ms time bins (we chose 10 ms because it is the membrane time constant of the neuron used later in the simulations), and demonstrates that nothing characterizes the periods when the pattern is present. The right panel plots the individual firing rates averaged over the whole period. Neurons involved in the pattern are shown in red. Again, nothing characterizes them in terms of firing rates. Detecting the pattern thus requires taking the spike times into account.
doi:10.1371/journal.pone.0001377.g001

realistically, pattern presentations occur at unpredictable times, separated by long 'distractor' periods and if the pattern does not involve all the afferents? Does it make sense to use the beginning of the pattern as a time reference, and does the postsynaptic spike latency with respect to this reference still decrease?

To answer these questions we inserted an arbitrary pattern at various times into randomly generated 'distractor' spike trains, as in Fig 1, and investigated whether a single receiving STDP neuron, with a 10 ms membrane time constant, was able to learn it in an unsupervised manner. To be precise, we simulated a

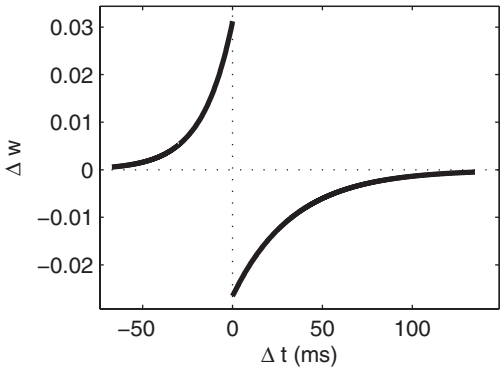


Figure 2. The STDP modification function. We plotted the additive weight updates as a function of the difference between the presynaptic spike time and the postsynaptic one. We used an exponential law (see Materials and Methods). The left part corresponds to Long Term Potentiation (LTP) and the right part to Long Term Depression (LTD).
doi:10.1371/journal.pone.0001377.g002

population of 2,000 afferents firing continuously for 450 s (see Materials and Methods for details). Most of the time (3/4 of the time in the baseline simulation) the afferents fired according to a Poisson process with variable instantaneous firing rates. Spiking activity in the brain is usually assumed to follow roughly Poisson statistics, hence this choice, but here it is not crucial: what matters is that the afferents fire stochastically and independently. But every now and then, at random times, half of these afferents left the stochastic mode for 50 ms and adopted a precise firing pattern. This repeated pattern had roughly the same spike density as the stochastic distractor part, so as to make it invisible in terms of firing rates. To be precise the firing rate averaged over the population and estimated over 10 ms time bins has a mean of 64 Hz and a standard deviation of less than 2 Hz (this firing rate is even more constant than in the 100 afferent case of Fig. 1 because of the law of large numbers). We further increased the difficulty by adding a permanent 10 Hz Poissonian spontaneous activity to all the neurons, and by adding a 1 ms jitter to the pattern. Intriguingly, we will see that one single Leaky Integrate-and-Fire (LIF) neuron receiving inputs from all the afferents, acting as a coincidence detector (see Fig. 3), and implementing STDP, is perfectly able to solve the problem and learns to respond selectively to the start of the repeating pattern.

RESULTS

At the beginning of a first simulation the 2,000 synaptic weights are all equal to 0.475 (arbitrary units normalized in the range [0,1]). The neuron is therefore non-selective. Since the presynaptic spike density – on its 10 ms time scale – is almost constant, it discharges periodically (see Fig. 4a). The greater are the initial

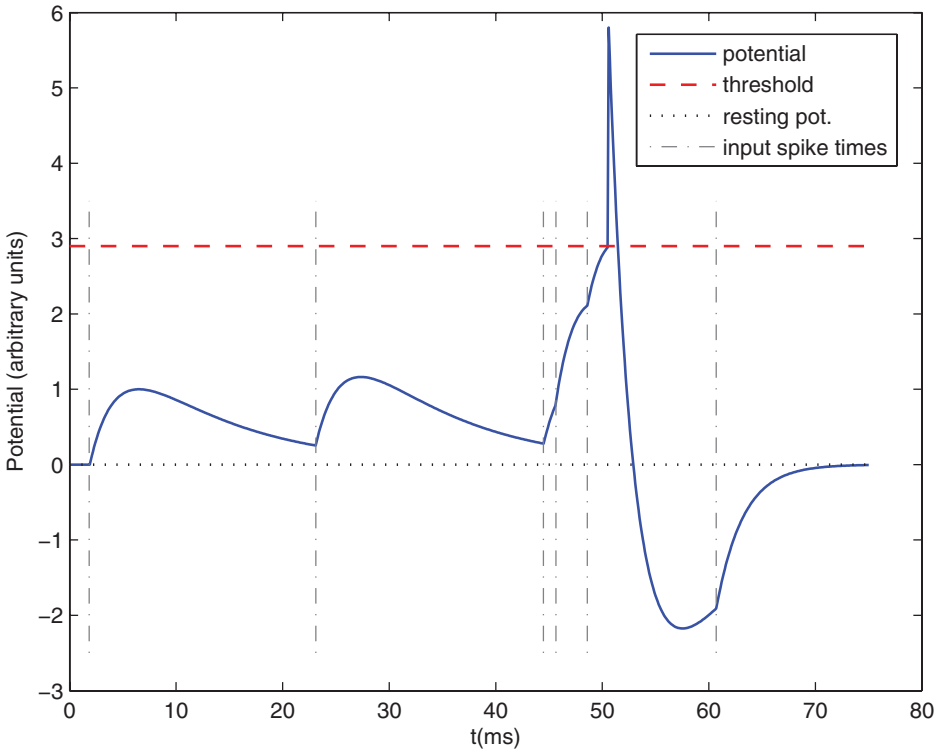


Figure 3. Leaky Integrate-and-Fire (LIF) neuron. Here is an illustrative example with only 6 input spikes. The graph plots the membrane potential as a function of time, and clearly demonstrates the effects of the 6 corresponding Excitatory PostSynaptic Potentials (EPSP). Because of the leak, for the threshold to be reached the input spikes need to be nearly synchronous. The LIF neuron is thus acting as a coincidence detector. When the threshold is reached, a postsynaptic spike is fired. This is followed by a refractory period of 1 ms and a negative spike-afterpotential.
doi:10.1371/journal.pone.0001377.g003

weights (or the lower the threshold), the smaller is the period (here it is about 16 ms, the initial firing rate is thus about 63 Hz). Each time a discharge occurs we update the synaptic weights using the STDP rule of Fig. 2, and clip them in the range [0,1]. At this stage, the neuron discharges both outside and inside the pattern (represented by grey rectangles on Fig. 4). In the first case presynaptic and postsynaptic spike times are uncorrelated, and since $a^-\tau^- > a^+\tau^+$ (where a^- and τ^- are respectively the LTD learning rate and time constant, and a^+ and τ^+ are the same parameters for LTP, see Materials and Methods), STDP leads to an overall weakening of synapses[15] (note: if no repeating patterns were inserted STDP would thus gradually decrease the synaptic weights until the threshold would not be reached any longer). But in the second case, by reinforcing the synaptic connections with the afferents that took part in firing the neuron, STDP increases the probability that the neuron fires again next time the pattern is presented (reinforcement of causality link). As a result, selectivity to the pattern emerges, here after about 13.5 s (see Fig. 4b) that is after only about 70 pattern presentations and 700 discharges: the neuron gradually stops discharging outside the pattern (no false alarms), while it does discharge most of the time when the pattern is presented (high hit rate), and can even fire twice per pattern as in the case illustrated here. Chance determines which part(s) of the pattern the neuron becomes selective to at this stage (i.e. the postsynaptic spike latency(ies), with respect to the beginning of the pattern here about 5 ms and 40 ms). However the increase in selectivity usually rapidly leads to only one discharge per pattern, here at about 40 ms.

Once selectivity to the pattern has emerged STDP has another major effect. Each time the neuron discharges in the pattern, it reinforces the connections with the presynaptic neurons that fired slightly before in the pattern. As a result next time the pattern is presented the neuron is not only more likely to discharge to it, but it will also tend to discharge earlier. In other words, the postsynaptic spike latency locks itself to the pattern and decreases steadily (with respect to the beginning of the pattern). However, it cannot decrease endlessly. There is a convergence by saturation when all the spikes in the pattern that precede the postsynaptic spike already correspond to maximally potentiated synapses, and all are necessary to reach the threshold. This usually occurs when the latency is already very short, the value depending on the threshold, although it could occur even earlier if the pattern has a zone with low spike density. Spikes outside the pattern cannot contribute efficiently to the membrane potential: since their times are stochastic, STDP usually depresses the corresponding synapses. We end up with a bimodal weight distribution with synapses either maximally potentiated or fully depressed (as predicted by van Rossum et al[24]).

Here this convergence occurs after about 2000 discharges. At this stage, the postsynaptic spike latency (with respect to the beginning of the pattern) is about 4 ms (see Fig. 4c). After convergence the hit rate is then 99.1% with no false alarms (estimated on the last 150 s). Notice that the signal/noise ratio has increased with respect to the situation in Fig. 4b, that is the potential reached on distractor periods is farther from the threshold. Among the 2,000 synapses, 383 are fully potentiated (weight$\approx$1), while the rest of them are almost completely depressed (weight$\approx$0). All of the potentiated synapses correspond to afferents involved in the pattern. The fact that there is no false alarms means once the learning has been done, a neuron just waits for its

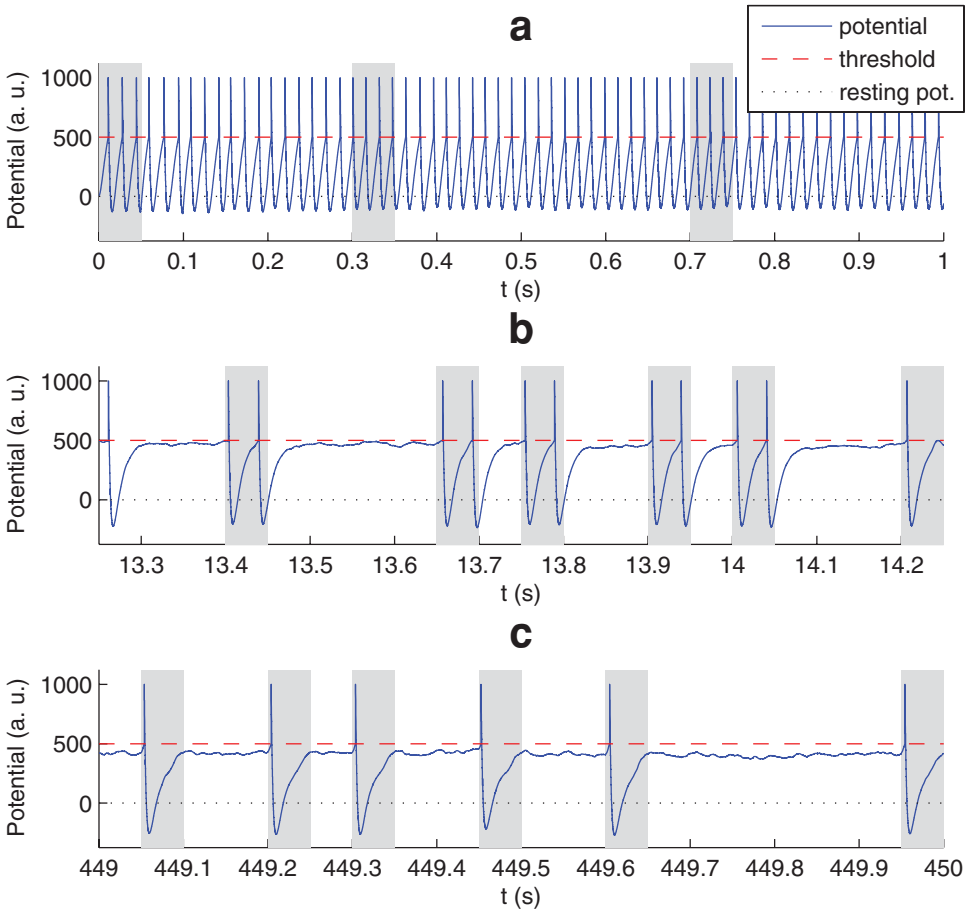


Figure 4. Overview of the 450 s simulation. Here we plotted the membrane potential as a function of simulation time, at the beginning, middle, and end of the simulation. Grey rectangles indicate pattern presentations. (a) At the beginning of the simulation the neuron is non-selective because the synaptic weights are all equal. It thus fires periodically, both inside and outside the pattern. (b) At t≈13.5 s, after about 70 pattern presentations and 700 discharges, selectivity to the pattern is emerging: gradually the neuron almost stops discharging outside the pattern (no false alarms), while it does discharge most of the time the pattern is present (high hit rate), here even twice (c) End of the simulation. The system has converged (by saturation). Postsynaptic spike latency is about 4 ms. Hit rate is 99.1% with no false alarms (estimated on the last 150 s).
doi:10.1371/journal.pone.0001377.g004

preferred stimulus, and need never forget what it has learned. The model thus predicts that fully specified neurons might actually have very low spontaneous rates, whereas higher rates might characterize less well specified cells.

Fig. 5 shows the latency reduction (with respect to the beginning of the pattern) during the learning stage until it stabilizes at a minimum of about 4 ms. Apart from the initial part (before selectivity emerges) the curve looks similar to those observed in earlier work with discrete spike volleys[17]. By convention the latency is 0 when the neuron discharged outside the pattern, that is when it generated a false alarm. There are no false alarms after the 676th discharge, that is for the last 436 s of simulation.

Fig. 6 illustrates the situation after convergence. It can be seen that STDP has potentiated most of the synapses that correspond to the earliest spikes of the pattern (Fig. 6a), and depressed most of the synapses that correspond to presynaptic spikes which follow the postsynaptic one, as in the previous work with discrete volleys [15,17,18]. This results in a sudden increase in membrane potential when the neuron starts integrating the pattern, and the threshold is quickly reached (Fig. 6b). Notice that all the synaptic connections with afferents not involved in the pattern have been completely depressed.

We performed 100 similar simulations with different pseudo-randomly generated spike trains and patterns. Our criteria for a 'successful' simulation were: convergence to a state with a postsynaptic latency inferior to 10 ms, a hit rate superior to 98% and no false alarms. This occurred in 96% of the cases. For the remaining 4%, the neurons stopped firing when too many discharges occurred outside the pattern in a row (leading to an overall weakening of synapses, so the threshold was no longer reached).

We ran other batches of 100 simulations to systematically investigate the impact on this 96% success performance of five parameters.

The first one is the pattern relative frequency (i.e sum of pattern durations over total duration ratio, assuming a fixed pattern duration of 50 ms), 1/4 in the baseline condition, and Fig. 7a shows its effect. We see that while the performance is very high as long as the ratio is above 15%, with smaller values the probability of success drops. This means the pattern needs to be consistently present for the STDP to learn it. However, this applies only at the beginning (say during the first 1000 discharges). Here we used a constant pattern frequency, but after the initial part the neuron has already become selective to the pattern, so presenting longer

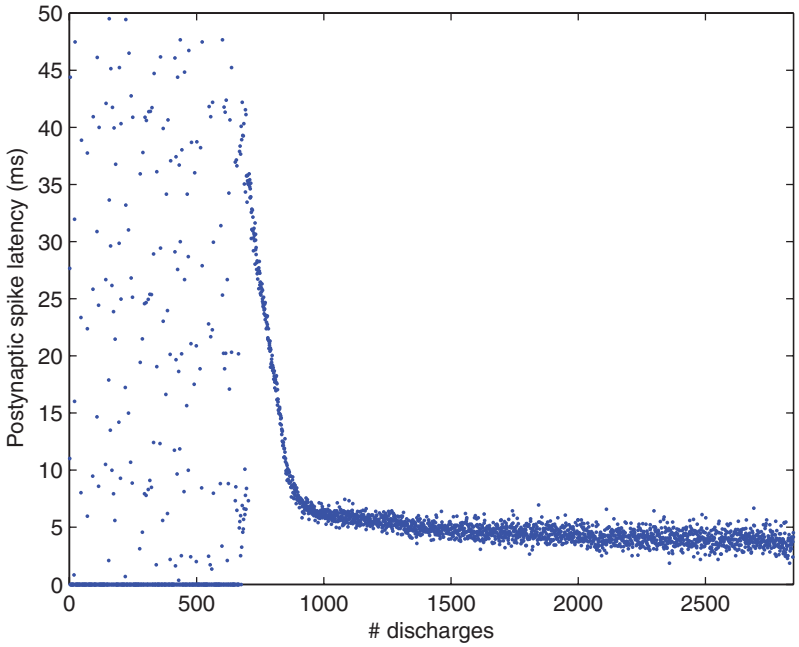


Figure 5. Latency reduction. Here we plotted the postsynaptic latency as a function of the number of discharges (by convention the latency is 0 when the neuron discharged outside the pattern, i.e. when it generated a false alarm). We clearly distinguish 3 periods: the beginning, when the neuron is non-selective; the middle, when selectivity has emerged and STDP is 'tracking back' through the pattern; and the end, when the system has converged towards a fast and reliable pattern detector.
doi:10.1371/journal.pone.0001377.g005

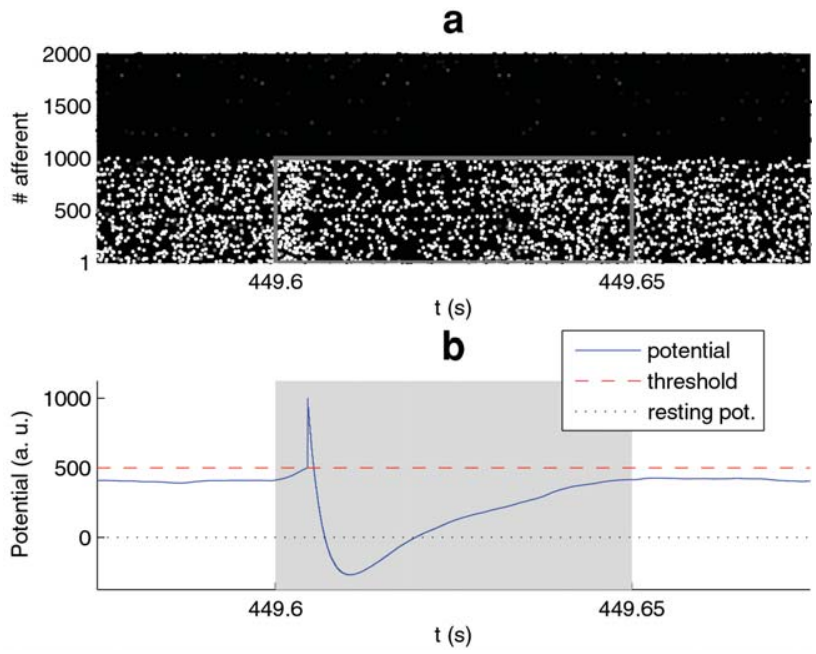


Figure 6. Converged state (a) we represented the spike trains of the 2,000 afferents. We have reordered the afferents with respect to Fig. 1 so that afferents 1–1000 are involved in the pattern, and afferents 1001–2000 are not and we use a color code ranging from black for spikes that correspond to completely depressed synapses (weight = 0) to white for spikes that correspond to maximally potentiated synapses (weight = 1). This allows the visualization of the spikes which generate a significant EPSP and those which do not. The pattern is represented with a grey line rectangle. Notice the cluster of white spikes at the beginning of it: STDP has potentiated most of the synapses that correspond to the earliest spikes of the pattern. Note that virtually all the synaptic connections with afferents not involved in the pattern have been completely depressed. (b) The membrane potential is plotted as a function of time, over the same range as above. We clearly see the sudden increase that corresponds to the above-mentioned cluster.
doi:10.1371/journal.pone.0001377.g006

distractor periods does not perturb the learning at all. We also tried to change the pattern duration while maintaining its relative frequency at 1/4. It turns out that what makes the detection difficult is the delay between two pattern presentations, not the pattern duration itself. Since we kept the pattern relative frequency constant, this delay increased with the pattern duration so the performance dropped: 97% with a 40 ms pattern, 96% with 50 ms, 93% with 60 ms, 59% with 100 ms and 46% with 150 ms. However we think this delay is more naturally investigated by changing the pattern relative frequency as in Fig. 7a.

The second parameter we investigated is the amount of jitter (1 ms in the baseline condition), and Fig. 7b shows its influence. We see that the performance is very good for jitter levels lower than 3 ms. For larger amounts of jitter the spike coincidences are lost, and the STDP weight updates are inaccurate, so the learning is impaired. In the brain millisecond spiking precision has been reported in many structures, including the retina[25,26], the Lateral Geniculate Nucleus[27,28], the visual cortex[29,30], the somatosensory system[31,32] and the auditory system[33]. Some authors report higher variability, but this could result from non controlled variables rather than intrinsic noise (see Discussion).

The third parameter is the proportion of afferents involved in the pattern (1/2 in the baseline condition), and Fig. 7c shows its influence. The threshold was scaled proportionally. Not surprisingly, with fewer afferents involved in the pattern, it becomes harder to detect, but it is still detected more than half of the times when only 1/3 of the afferents are involved in the pattern. Note that the other 2/3 of afferents are discarded by STDP. This suggests that activity-driven mechanisms could select a small set of 'interesting' afferents among a much bigger set of initially connected afferents, probably specified genetically, a phenomenon known as 'developmental exuberance' for which there is considerable experimental evidence[34].

The fourth parameter is the initial weight (0.475 in the baseline condition) and Fig. 7d shows its influence. Recall discharges outside the pattern lead to an overall decrease of synaptic weights. If too many of them occur in a row the threshold may no longer be reachable. Thus a high initial value for the weights increases the resistance to discharges outside the pattern, leading to a better performance. High initial weights also cause the neuron to discharge at a high rate at the beginning of the learning process, when it is non-selective: 63 Hz for an initial weight of 0.475, 38 Hz for 0.325. These values may seem high in regard to usual experimental values. But first after only 13 s selectivity has emerged, and the neuron fires at a rate between 5 and 10 Hz. It is conceivable that electrophysiologists rarely record such short very active initial phases. Second, we consider here that the population of afferents is constantly firing with a mean rate of 64Hz. This is to make the problem of pattern detection harder, but if the afferents have less active periods, which is likely to occur in the brain, so will have the post-synaptic neuron. We also added Gaussian noise to the initial weights, with increasing standard deviation until 0.475 (thus equal to the mean). Following this noise addition the weights were clipped in [0,1]. This had no significant impact on the performance, at least in the present case when the initial weights are relatively high.

The fifth parameter is the proportion of missing spikes (0 in the base line condition). The threshold was scaled proportionally. Not surprisingly the number of successfully learned patterns decreases with the proportion of spikes deleted. However with a 10% deletion the pattern was correctly learnt 82% of the time, demonstrating that the system is quite robust to spike deletion.

We also tried changing the membrane time constant τ_m (10 ms in the baseline condition), scaling the threshold proportionally. This had little impact on the performance (79% success with $\tau_m = 5$ ms, 88% with $\tau_m = 20$ ms), but it did have an impact on the minimal latency that is reached after convergence. A smaller time constant (and the smaller threshold that goes with it) causes the neuron to be interested in more coincident spikes. The system converges when the very few nearly coincident first spikes of the pattern all correspond to maximally potentiated synapses, and the postsynaptic spikes is fired just after them. The final latency is thus shorter than the one we have with a longer time constant, which enables the neuron to integrate spikes over a longer time window.

Taken together these results demonstrate that the learning is amazingly robust to the model parameters. We thus believe that we have captured a mechanism than emerges from STDP rather than from a precise neural model configuration. While we admit it is still somewhat speculative to affirm that a similar mechanism takes place in the brain, it is at least very plausible.

DISCUSSION

Our first claim is that the main results previously obtained for STDP based learning with the highly simplified scheme of discrete spike volleys[15–18] still stand in this more challenging continuous framework. This means that global discontinuities such as saccades

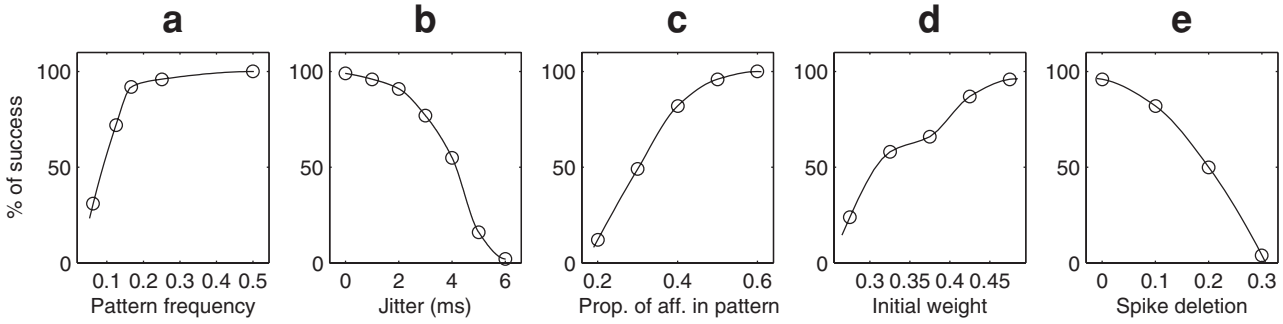


Figure 7. Resistance to degradations (100 trials). (a) Percentage of successful trials as a function of the pattern frequency (pattern duration/the total duration, given a fixed pattern length of 50 ms). The pattern needs to be consistently present, at least at the beginning, for the STDP to start the learning process. (b) Percentage of successful trials as a function of jitter. For jitter greater than 3 ms spike coincidences are lost and the STDP weight updates are inaccurate, so the learning is impaired (c) Percentage of successful trials as a function of the proportion of afferents involved in the pattern. Performance is good if this proportion is above 1/3 (d) Percentage of successful trials as a function of the initial weights. With a high value the neuron can handle more discharges outside the pattern. (e) Percentage of successful trials as a function of the proportion of spikes deleted. With a 10% deletion the pattern was correctly learnt in 82% of the cases.
doi:10.1371/journal.pone.0001377.g007

or micro-saccades in vision and sniffs in olfaction[35], or brain oscillations in general[23] are not necessary for STDP-based learning of temporal patterns (although they will almost certainly help). Temporal code skeptics often point out the fact that neurons would need to know a time reference to decode a temporal code, and we see here that this is not necessary: as long as there are recurrent spike patterns in the inputs, and even if they are embedded in equally dense 'distractor' spike trains, a neuron equipped with STDP can potentially find them in only a few tens of pattern presentations, and will gradually respond faster and faster when the pattern is presented, by potentiating synapses that correspond to the earliest spikes of the patterns, and depressing all the others. This last point strongly reinforces the idea that a substantial amount of information could be available very rapidly, in the very first spikes evoked by a stimulus[36].

It is worth mentioning that the proposed learning scheme is fully unsupervised. No teaching signal tells the neuron when to learn nor labels the inputs. Biologically plausible mechanisms for supervised learning of spike patterns have also been proposed[37].

It is also surprising to see how such a simple mechanism can solve a problem as complex as spike pattern detection. However, there is no consensus on the definition of a spike pattern, and we admit ours is quite simple: here a pattern is seen as a succession of coincidences. A Leaky Integrate and Fire (LIF) neuron is known to be capable of coincidence detection, and it has even been proposed that this is its main function in the brain[38,39]. Here the membrane time constant (10 ms) is shorter than the duration of the pattern (50 ms), and so the LIF neuron can never be selective to the whole pattern. Instead, it is selective to 'one coincidence' of the pattern at a time, that is, selective to the nearly simultaneous arrival of certain spikes, just as it occurs in one subdivision of the pattern. At the beginning of the learning process STDP will cause the LIF neuron to become selective to one such coincidence (chance determines which one). Then STDP will track back through the pattern, from one coincidence to the previous one, until the initial coincidence is reached and the chain of causality is stopped. At this point the neuron is selective only to the simultaneous arrival of the pattern's earliest spikes, and can serve as 'earliest predictor' of the subsequent spike events[15,16,19], at the risk of triggering a false alarm if these subsequent events don't occur, but with the benefit of being very reactive.

This contrasts with approaches where the whole pattern needs to be taken into account, sometimes including finer structural aspects such as spike orders or relative delays[2,3,40,41]. But neuronal mechanisms able to reliably decode such structures have to be proposed and looked for in the brain. One appealing candidate mechanism is the synfire chain[42] but direct evidence for their existence is still fairly limited[43]. Here we limit the notion of pattern to successive coincidences, and suggest a way such patterns could be decoded, using widely accepted neurophysiological mechanisms, namely coincidence detection and STDP.

Another limitation of this work is the excitatory-only scheme. Consequently, something like 'afferent A must not spike' cannot be learnt, only 'positive patterns' can. However, evidence for plasticity in inhibitory synapses in the brain is weak and inhibition is often assumed to be non-selective. So we propose that most of the selectivity could be achieved using only excitatory synapses, as in this model.

Whether spike times contain additional information with respect to discharge rates has been the object of an ongoing debate for some time. Electrophysiologists have tried to answer this question mostly by recording neurons in sensory and motor systems with a repeating stimulus or action, and looking at inter-trial variability of the spike times. Some claim that spike times can be very reliable while others are more skeptical (see ref [22,44] for reviews). Given that the simple and cheap mechanism exposed here reliably detects spatio-temporal spike patterns, it is hard to believe that the brain did not evolve to use at least the form of temporal coding exposed above ('successive coincidences'), unless there is an unavoidable intrinsic source of noise in the integrate-and-fire mechanism that makes all spike times unreliable. The main source for this sort of noise is probably at the level of synaptic transmission[45], since neurons stimulated directly by current injection in the absence of synaptic input give highly stereotyped and precise responses[46]. However, spike times can be very reliable in some experiments[22,44], particularly in the auditory cortex, proving that reliable synapses do exist. So we argue that variability in other recorded spike times, in particular in the visual system, could come from non-controlled variables that might also affect neuronal activation, such as attention, eye movements, mental imagery, top-down effects etc. As Barlow wrote about neural responses in 1972, ''their apparently erratic behavior was caused by our ignorance, not the neuron's incompetence.''[47]

We would like to emphasize the fact that the approach presented here is generic. It is not limited to sensory systems, and it could be applied to either experimental or model-generated data. The first step would be to see if STDP finds spike patterns in the data. Providing it does, the second step would be to understand what those patterns mean by solving the corresponding inverse problem.

What happens if there is more than one repeating pattern present in the input? We verified that as the learning progresses, the increasing selectivity of the postsynaptic neuron rapidly prevents it from responding to several patterns. Instead, it picks one (chance determines which one), and becomes selective to it and only to it. To learn the other patterns other neurons are needed.

A competitive mechanism could ensure they optimally cover all the different patterns and avoid learning the same ones. Such a mechanism could be implemented through inhibitory horizontal connections between neurons, such that as soon as one neuron fires, it could prevent other cells from learning the same pattern, as in previous work[48]. The neural population would then self-organize to cover all the input patterns. The 'coverage' could be optimized using neurons that differ in their parameters (for example their thresholds), leading to more robust learning and detection. Furthermore a long input pattern can be coded by the successive firings of several STDP neurons, each selective to a different part of the pattern, and competition would prevent them all from tracking back through the pattern and clustering at the beginning. Note that within such a competitive framework a pattern detection probability of 50% is hardly a disaster: it means that with 2 neurons the risk that one pattern is not detected is 25%, with 3 neurons 12.5%, with 4 neurons 6.25% and so on. The system could then work with suboptimal parameters (highlighted in Fig. 7), for example weaker initial weights.

Further work is needed to evaluate this form of competitive network. However in this paper we wanted to stress the fact that *one* single LIF neuron equipped with STDP is consistently able to detect *one* arbitrary repeating spatio-temporal spike pattern embedded in equally dense 'distractor' spike trains, which is a remarkable demonstration of the potential for such a scheme.

MATERIALS AND METHODS

The simulations were performed using MATLAB R14 (Mathworks 2005, Natick MA). The source code is available from the authors upon request.

Poisson spike trains

The spike trains were prepared before the simulation (Fig. 1 illustrates the type of spike trains we used, though with a smaller set of neurons). For memory issues instead of using spike trains defined over a 450 seconds period, we pasted the same 150s long pattern three times (this repetition had no impact on the results). Each afferent emits spikes independently using a Poisson process with a variable instantaneous firing rate r, that varies randomly between 0 and 90 Hz. The maximal rate change s was chosen so that the neuron could go from 0 to 90 Hz in 50 ms. To be precise, time was discretized using a time step dt of 1 ms. At each time step:

1. the afferent has a probability of r.dt of emitting a spike (whose exact date is then picked randomly in the 1 ms time bin)
2. its instantaneous firing rate is modified: dr = s.dt where s is the speed of rate change (in Hz/s), and clipped in [0, 90] Hz.
3. its speed of rate change is modified by ds, randomly picked from a uniform distribution over [−360+360] Hz/s, and clipped in [−1800+1800] Hz/s

Note that we chose to apply the random change to s as opposed to r so as to have a continuous s function and a smoother r function.

As mentioned in the Discussion, a limitation of this work is the excitatory-only scheme. Consequently, something like 'afferent A must not spike' cannot be learnt, only 'positive patterns' can. We thus wanted a pattern in which all the afferents spike at least once. We could have made up such a pattern, but we wanted the pattern to have exactly the same statistics as the Poisson distractor part (to make the pattern detection harder), so we preferred to randomly pick a 50 ms period of the original Poisson spike trains and to 'copy-paste' it (see below). To make sure this randomly selected period did contain a spike from each afferent we implemented a mechanism that triggers a spike whenever an afferent has been silent for more than 50 ms (leading to a minimal firing rate of 20 Hz). Clearly, such mechanism is NOT implemented in the brain. It is just an artifice we used here to make the pattern detection harder. As a result the average firing rate was 54 Hz, and not the 45 Hz we would have without this additional mechanism.

Once the random spike train has been generated, a part of it, defined as the 'pattern' to be repeated, is 'copy-pasted'. This 'copy-paste' does not involve the last 1000 afferents (obviously the indices are arbitrary), which conserve their original spike trains. But we discretize the spike trains of the first 1000 afferents into 50 ms sections. We randomly pick one of these sections and copy the corresponding spikes. Then we randomly pick a certain number of these sections (1/4 in the baseline condition), avoiding consecutive ones, and replace the original spikes by the copied ones. A jitter was added before the pasting operation, picked from a Gaussian distribution with mean zero and standard deviation 1 ms (in the baseline condition).

After this 'copy-paste' operation a 10 Hz Poissonian spontaneous activity was added, to all neurons and all the time. The total activity was thus 64 Hz on average, and spontaneous activity represented about 16% of it.

Leaky Integrate and Fire (LIF) neuron (see Fig. 3)

For computational reasons we modeled the LIF neuron using Gerstner's Spike Response Model (SRM)[16,49]. That is instead of solving the membrane potential differential equation we used kernels to model the effect of presynaptic and postsynaptic spikes on the membrane potential. Each presynaptic spike j, with arrival time t_j, is supposed to add to the membrane potential an Excitatory Post-Synaptic Potential (EPSP) of the form:

$$\varepsilon(t-t_j) = K\cdot\left(\exp\left(-\frac{t-t_j}{\tau_m}\right)-\exp\left(-\frac{t-t_j}{\tau_s}\right)\right)\cdot\Theta(t-t_j)$$

where τ_m is the membrane time constant (here 10 ms), τ_s is the synapse time constant (here 2.5 ms), Θ is the Heavyside step function:

$$\Theta(s) = \begin{cases} 1 & if \quad s\geq 0 \\ 0 & if \quad s<0 \end{cases}$$

and K is just a multiplicative constant chosen so that the maximum value of the kernel is 1 (the voltage scale is arbitrary in this paper).

The last emitted postsynaptic spike i has an effect on the membrane potential modeled as follows:

$$\eta(t-t_i) = \mathrm{T}\cdot\left(K_1\cdot\exp\left(-\frac{t-t_i}{\tau_m}\right)-K_2\cdot\left(\exp\left(-\frac{t-t_i}{\tau_m}\right)-\exp\left(-\frac{t-t_i}{\tau_s}\right)\right)\right)\cdot\Theta(t-t_i)$$

where T is the threshold of the neuron (here 500, arbitrary units). The first term models the positive pulse and the second one the negative spike-afterpotential that follows the pulse (see Fig. 3). Here we used $K_1 = 2$ and $K_2 = 4$. For simplicity, the resting potential is supposed to be zero, but a non zero value would simply shift the kernel, and shifting the threshold by the same value would lead to the same computation.

Both ε and η kernels were rounded to zero when respectively $t-t_j$ and $t-t_i$ were greater than $7\cdot\tau_m$.

At any time the membrane potential is:

$$p = \eta(t-t_i) + \sum_{j/t_j>t_i} w_j\cdot\varepsilon(t-t_j)$$

where the w_j are the excitatory synaptic weights, between 0 and 1 (arbitrary units).

This SRM formulation allows us to use event-driven programming: we only compute the potential when a new presynaptic spike is integrated. We then estimate numerically if the corresponding EPSP will cause the threshold to be reached in the future and at what date. If it is the case, a postsynaptic spike is scheduled. Such postsynaptic spike events cause all the EPSPs to be flushed, and a new t_i is used for the η kernel. There is then a refractory period of 1 ms, during which the neuron is not allowed to fire.

Spike Timing Dependent Plasticity

An exponential update rule (see Fig. 2):

$$\Delta w_j = \begin{cases} a^+\cdot\exp\left(\frac{t_j-t_i}{\tau^+}\right) & \text{if} \quad t_j\leq t_i \quad \text{(LTP)} \\ -a^-\cdot\exp\left(-\frac{t_j-t_i}{\tau^-}\right) & \text{if} \quad t_j>t_i \quad \text{(LTD)} \end{cases}$$

with the time constants $\tau^+ = 16.8$ ms and $\tau^- = 33.7$ ms, provides a reasonable approximation of the synaptic modification observed experimentally[13].We restricted the learning window to $[t_i-7\cdot\tau^+, t_i]$ for LTP and to $[t_i, t_i+7\cdot\tau^-]$ for LTD. For each afferent, we also limited LTP (respectively LTD) to the last (first)

presynaptic spike before (after) the postsynaptic one ('nearest spike' approximation). We did not take the effects of finer triplet of spikes[50] into account.

It was found that small learning rates led to more robust learning. We used $a^+ = 0.03125$ and $a^- = 0.85 \cdot a^+$ Following learning the weights were clipped to [0,1]. Note that all synapses remain excitatory: there is no inhibition in all these simulations.

ACKNOWLEDGMENTS

We thank Rufin VanRullen for reading the manuscript and making pertinent comments and Pierre Bayerl for his participation in an earlier stage of this project.

Author Contributions

Conceived and designed the experiments: ST TM RG. Performed the experiments: TM. Analyzed the data: TM. Wrote the paper: ST TM RG.

REFERENCES

1. Frostig RD, Frysinger RC, Harper RM (1990) Recurring discharge patterns in multiple spike trains. II. Application in forebrain areas related to cardiac and respiratory control during different sleep-waking states. Biol Cybern 62: 495–502.
2. Prut Y, Vaadia E, Bergman H, Haalman I, Slovin H, et al. (1998) Spatiotemporal structure of cortical activity: properties and behavioral relevance. J Neurophysiol 79: 2857–2874.
3. Fellous JM, Tiesinga PH, Thomas PJ, Sejnowski TJ (2004) Discovering spike patterns in neuronal responses. J Neurosci 24: 2989–3001.
4. Markram H, Lubke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275: 213–215.
5. Bi GQ, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18: 10464–10472.
6. Zhang LI, Tao HW, Holt CE, Harris WA, Poo M (1998) A critical window for cooperation and competition among developing retinotectal synapses. Nature 395: 37–44.
7. Feldman DE (2000) Timing-based LTP and LTD at vertical inputs to layer II / III pyramidal cells in rat barrel cortex. Neuron 27: 45–56.
8. Vislay-Meltzer RL, Kampff AR, Engert F (2006) Spatiotemporal specificity of neuronal activity directs the modification of receptive fields in the developing retinotectal system. Neuron 50: 101–114.
9. Mu Y, Poo MM (2006) Spike Timing-Dependent LTP/LTD Mediates Visual Experience-Dependent Plasticity in a Developing Retinotectal System. Neuron 50: 115–125.
10. Cassenaer S, Laurent G (2007) Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. Nature.
11. Meliza CD, Dan Y (2006) Receptive-field modification in rat visual cortex induced by paired visual stimulation and single-cell spiking. Neuron 49: 183–189.
12. Jacob V, Brasier DJ, Erchova I, Feldman D, Shulz DE (2007) Spike Timing-Dependent Synaptic Depression in the *In Vivo* Barrel Cortex of the Rat. The Journal of Neuroscience 27: 1271–1284.
13. Bi GQ, Poo MM (2001) Synaptic modification by correlated activity : Hebb's postulate revisited. Ann Rev Neurosci 24: 139–166.
14. Young JM, Waleszczyk WJ, Wang C, Calford MB, Dreher B, et al. (2007) Cortical reorganization consistent with spike timing–but not correlation-dependent plasticity. Nat Neurosc 10: 887–895.
15. Song S, Miller KD, Abbott LF (2000) Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3: 919–926.
16. Gerstner W, Kistler WM (2002) Spiking neuron models. Cambridge University Press.
17. Guyonneau R, VanRullen R, Thorpe SJ (2005) Neurons tune to the earliest spikes through STDP. Neural Comput 17: 859–879.
18. Masquelier T, Thorpe S (2007) Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. PLoS Comput Biol 3.
19. Mehta MR, Quirk MC, Wilson MA (2000) Experience-dependent asymmetric shape of hippocampal receptive fields [see comments]. Neuron 25: 707–715.
20. Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383: 76–81.
21. Hopfield JJ (1995) Pattern recognition computation using action potential timing for stimulus representation. Nature 376: 33–36.
22. VanRullen R, Guyonneau R, Thorpe SJ (2005) Spike times make sense. Trends Neurosci 28: 1–4.
23. Fries P, Nikolic D, Singer W (2007) The gamma cycle. Trends Neurosci 30: 309–316.
24. VanRossum MCW, Bi GQ, Turrigiano GG (2000) Stable Hebbian Learning from Spike Timing-Dependent Plasticity. The Journal of Neuroscience 20: 8812–8821.
25. Berry MJ 2nd, Meister M (1998) Refractoriness and neural precision. J Neurosci 18: 2200–2211.
26. Uzzell VJ, Chichilnisky EJ (2004) Precision of spike trains in primate retinal ganglion cells. J Neurophysiol 92: 780–789.
27. Reinagel P, Reid RC (2000) Temporal coding of visual information in the thalamus. J Neurosci 20: 5392–5400.
28. Liu RC, Tzonev S, Rebrik S, Miller KD (2001) Variability and information in a neural code of the cat lateral geniculate nucleus. J Neurophysiol 86: 2789–2806.
29. Bair W, Koch C (1996) Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. Neural Comput 8: 1185–1202.
30. Buracas GT, Zador AM, DeWeese MR, Albright TD (1998) Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron 20: 959–969.
31. Johansson RS, Birznieks I (2004) First spikes in ensembles of human tactile afferents code complex spatial fingertip events. Nat Neurosci 7: 170–177.
32. Boloori AR, Stanley GB (2006) The dynamics of spatiotemporal response integration in the somatosensory cortex of the vibrissa system. J Neurosci 26: 3767–3782.
33. Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature 426: 442–446.
34. Innocenti GM, Price DJ (2005) Exuberance in the development of cortical networks. Nat Rev Neurosci 6: 955–965.
35. Uchida N, Kepecs A, Mainen ZF (2006) Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. Nat Rev Neurosci 7: 485–491.
36. VanRullen R, Thorpe SJ (2001) Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. Neural Comput 13: 1255–1283.
37. Gutig R, Sompolinsky H (2006) The tempotron: a neuron that learns spike timing-based decisions. Nat Neurosci 9: 420–428.
38. Abeles M (1982) Role of the cortical neuron: integrator or coincidence detector? Isr J Med Sci 18: 83–92.
39. Konig P, Engel AK, Singer W (1996) Integrator or coincidence detector? The role of the cortical neuron revisited. Trends Neurosci 19: 130–137.
40. Frostig RD, Frostig Z, Harper RM (1990) Recurring discharge patterns in multiple spike trains. I. Detection. Biol Cybern 62: 487–493.
41. Abeles M, Gat I (2001) Detecting precise firing sequences in experimental data. J Neurosci Methods 107: 141–154.
42. Abeles M (1991) Corticonics: neural circuits of the cerebral cortex. Cambridge; New York: Cambridge University Press. pp xiv, 280.
43. Abeles M (2004) Neuroscience. Time is precious. Science 304: 523–524.
44. Stein RB, Gossen ER, Jones KE (2005) Neuronal variability: noise or part of the signal? Nat Rev Neurosci 6: 389–397.
45. Movshon JA (2000) Reliability of neuronal responses. Neuron 27: 412–414.
46. Mainen ZF, Sejnowski TJ (1995) Reliability of Spike Timing in Neocortical Neurons. Science 268: 1503–1506.
47. Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1: 371–394.
48. Guyonneau R, VanRullen R, Thorpe SJ (2004) Temporal codes and sparse representations: a key to understanding rapid processing in the visual system. J Physiol Paris 98: 487–497.
49. Gerstner W (1995) Time structure of the activity in neural network models. Phys Rev E 51: 738–758.
50. Pfister J, Gerstner W (2006) Triplets of Spikes in a Model of Spike Timing-Dependent Plasticity. The Journal of Neuroscience 26: 9673–9682.

Appendix B

Conference abstracts & posters

B.1 Ultra-rapid visual form analysis using feedforward processing

Masquelier T, Guyonneau R, Guilbaud N, Allegraud J-M, Thorpe S J, 2005, "Ultra-rapid visual form analysis using feedforward processing" *Perception* 34 ECVP Abstract Supplement. `http://www.perceptionweb.com/abstract.cgi?id=v050563`

The speed with which humans and monkeys can detect the presence of animals in complex natural scenes constitutes a major challenge for models of visual processing. Here, we use simulations using SpikeNet (`http://www.spikenet-technology.com`) to demonstrate that even complex visual forms can be detected and localised with a feedforward processing architecture that uses the order of firing in a single wave of spikes to code information about the stimulus. Neurons in later recognition layers learn to recognise particular visual forms within their receptive field by increasing the synaptic weights of inputs that fire early in response to a stimulus. This concentration of weights on early firing inputs is a natural consequence of spike-time-dependent plasticity (STDP) (see Guyonneau et al. (2005)). The resulting connectivity patterns produce neurons that respond selectively to arbitrary visual forms while retaining a remarkable degree of invariance in image transformations. For example, selective responses are obtained with image size changes of roughly $\pm 20\%$, rotations of around $\pm 12^\circ$, and viewing angle variations of approximately $\pm 30^\circ$. Furthermore,there is also very good tolerance to variations in contrast and luminance and to the addition of noise or blurring. The performance of this neurally inspired architecture raises the possibility that our ability to detect animals and other complex

forms in natural scenes could depend on the existence of very large numbers of neurons in higher-order visual areas that have learned to respond to a wide range of image fragments, each of which is diagnostic for the presence of an animal part. The outputs of such a system could be used to trigger rapid behavioural responses, but could also be used to initiate complex and time-consuming processes that include scene segmentation, something that is not achieved during the initial feedforward pass.

B.2 Face feature learning with Spike Timing Dependent Plasticity

I presented preliminary results on STDP-based visual feature learning at the conference NeuroComp 06, Pont-à-Mousson, France.

The model I used is a simplified version of the one presented in Chapter 2. There are only two layers: the first one mimics V1 simple cells with a time-to-first-spike coding (this layer corresponds to S_1 in the complete model), and the subsequent one implements STDP (S_2 in the complete model). There are no complex cells, nor classification layers.

I applied it on face images, and STDP did extract face features.

B.2.1 Paper

FACE FEATURE LEARNING WITH SPIKE TIMING DEPENDENT PLASTICITY

Timothée Masquelier and Simon J Thorpe
Centre de Recherche Cerveau et Cognition
UMR 5549 CNRS - Université Paul Sabatier Toulouse 3
and
SpikeNetTechnology SARL, Labège, France
email: timothee.masquelier, simon.thorpe @cerco.ups-tlse.fr

ABSTRACT
Spike Timing Dependent Plasticity (STDP) is a learning rule that modifies synaptic strength as a function of the relative timing of pre-and postsynaptic spikes. Here we use this learning rule with neurons integrating spike trains coming from V1 orientation selective cells. Presenting natural images containing faces we observe that the neurons develop selectivity to face features. These results suggest that temporal codes may be a key to understanding the phenomenal processing speed achieved by the visual system, and argue that STDP can lead to fast and selective responses.

KEY WORDS
STDP, spiking neurons, learning, visual features, face recognition

1 Introduction

Temporal constraints pose a major challenge to models of face recognition in cortex. When two images are simultaneously flashed to the left and right of fixation, human subjects can make reliable saccades to the side where there is a face in as little as 100-110 ms[16]. If we allow 20-30 ms for motor delays in the oculomotor system, this implies that the underlying visual processing can be done in 80-90 ms. In monkeys, recent recordings from inferotemporal cortex (IT) shows that spike counts over time bins as small as 12.5 ms (that produce essentially a binary vector with either ones or zeros) and only about 100 ms after stimulus onset contain remarkably accurate information about the nature of a visual stimulus[10]. This sort of rapid processing presumably depends on the ability of the visual system to learn to recognize familiar visual forms. Quite how this learning occurs constitutes a major challenge for theoretical neuroscience.

Here we explored the capacity of a simple 2-layer feedforward network that has two key features. First, when stimulated with a flashed visual stimulus, the neurons in the first layer of the system (mimicking V1 orientation selective cells) fire asynchronously, with the most strongly activated neurons firing first. This mechanism has been shown to efficiently encode image information[21]. Second, neurons in the second layer implement Spike-Time Dependent Plasticity (STDP), which is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early[14, 9].

We demonstrate that when such a system is repeatedly presented with natural images containing faces, the second layer neurons will naturally become selective to patterns that are both salient and reliably present in the input, in this case face features. Furthermore, thanks to the use of competition, the neuron population self-organizes, each neuron learning a distinct pattern, so as to cover the whole variability of the inputs.

2 Model

2.1 V1 spike trains

When presented with an image, the first layer's cells, emulating V1 cells, detect edges with four preferred orientations and the more strongly a cell is activated the earlier it fires. This intensity-latency conversion is in accordance with recordings in V1 showing that response latency decreases with the stimulus contrast[1, 7] and with the proximity between the stimulus orientation and the cell's preferred orientation[4]. We also limit the number of spikes at this stage by introducing competition between V1 cells through a 1-Winner-Take-All mechanism: at a given location corresponding to one cortical column only the spike corresponding to the best matching orientation is propagated (sparsity is thus 25% at this stage). Note that k-Winner-Take-All mechanisms are easy to implement in the temporal domain using inhibitory GABA interneurons[15]. Those V1 spikes are then propagated asynchronously towards the second layer, where STDP is used.

2.2 STDP

STDP is a learning rule that modifies the strength of a neuron's synapses as a function of the precise temporal relations between pre- and post-synaptic spikes: an excitatory synapse receiving a spike before a postsynaptic one is emitted is potentiated (Long Term Potentiation) whereas its strength is weakened the other way around (Long Term

Depression)[12]. The amount of modification depends on the delay between these two events: maximal when pre- and post-synaptic spikes are close together, the effects gradually decrease and disappear with intervals in excess of a few tens of milliseconds[3, 22, 6]. Note that STDP is in agreement with Hebb's postulate because presynaptic neurons that fired slightly before the postsynaptic neuron are those which 'took part in firing it'. Several authors have studied the effect of STDP with Poisson spike trains[14, 19]. Here, we demonstrate the STDP's remarkable ability to detect statistical regularities in terms of earliest firing afferent patterns within visual spike trains, despite their very high dimensionality inherent to natural images.

3 An iterative process

Visual stimuli are presented sequentially and the resulting spike waves are propagated until the STDP layer. We process various scaled versions of the input image (with the same filter size). This results in cells with various receptive field sizes. It has been suggested that the mammalian visual system also performs multi-resolution processing using multiple receptive field sizes[5]. We use restricted receptive fields (*i.e*. cells only integrate spikes from a s x s square neighborhood in the V1 maps corresponding to one given processing scale) and weight sharing (*i.e*. each prototype cell is duplicated in retinotopic maps and at all scales). Starting with a random weight matrix (size = 4 x s x s) we present the first visual stimuli. Duplicated cells are all integrating the spike train and compete with each other. If no cell reaches its threshold nothing happens and we process the next image. Otherwise for each prototype the first duplicate to reach its threshold is the winner. A 1-Winner-Take-All mechanism prevents the other duplicated cells from firing. The winner thus fires and the STDP rule is triggered. Its weight matrix is updated, and the change in weights is duplicated at all positions and scales. This allows the system to learn patterns despite of changes in position and size in the training examples. We also use local inhibition between different prototype cells: when a cell fires at a given position and scale, it prevents all other cells from firing later at the same scale and within an s/2 x s/2 square neighborhood of the firing position. This competition prevents all the cells from learning the same pattern. Instead, the cell population self-organizes, each cell trying to learn a distinct pattern so as to cover the whole variability of the inputs.

If the stimuli have visual features in common (which should be the case if for example they contain similar objects), the STDP process will extract them. That is, for some cells we will observe convergence of the synaptic weights (by saturation), which end up being either maximally potentiated or fully depressed. During the convergence process synapses compete for control of the timing of postsynaptic spikes[14]. The winning synapses are those through which the earliest spikes arrive (on average)[14, 9], and this is true even in the presence of jitter and spontaneous activity[9]. This 'preference' for the earliest spikes is a key point since the earliest spikes, which correspond to the most salient regions of an image, have been shown to be the most informative[21]. During the learning the postsynaptic spike latency decreases[14, 9, 8]. After convergence, the responses become selective (in terms of latency)[9], here to visual features of intermediate complexity. Features can now be defined as clusters of afferents that are consistently among the earliest to fire. STDP detects these kinds of statistical regularities among the spike trains and creates one unit for each distinct pattern. This can be seen as a vector quantization process, which removes redundancy in the input and performs dimension reduction, thereby facilitating subsequent processing, for *e.g*. face recognition.

4 Results

We evaluated our STDP-based learning algorithm on a Caltech face dataset (available at www.vision.caltech.edu), using three prototype STDP cells. Faces are seen under various lighting condition and with very varied backgrounds.

Figure 1 illustrates the learning process (a video showing all the learning process (presentation by presentation) can be seen at http://cerco.ups-tlse.fr/∼masqueli/). Starting from random preferred stimuli, cells detect statistical regularities among the input spike trains after about one hundred discharges, and progressively develop selectivity to those patterns. A few hundred more discharges are needed to reach a stable state. Furthermore, the population of cells self-organizes: the blue cell learns the eyes and forehead, the green cell learns the nose and right eye, while the blue cell learns a coarse view of a whole face.

The background is generally not learned (at least not in priority), since backgrounds are almost always too different from one image to another for the STDP process to converge. Furthermore as STDP performs vector quantization from multiple examples as opposed to 'one shot learning', it will not learn the noise, nor anything too specific to a given example, with the result that it will tend to learn archetypical features.

Another key point is the natural trend of the algorithm to learn salient regions, simply because they correspond to the earliest spikes, with the result that neurons whose receptive fields cover salient regions are likely to reach their threshold (and trigger the STDP rule) before neurons 'looking' at other regions. This contrasts with more classical competitive learning approaches, where input normalization helps different input patterns to be equally effective in the learning process[13]. Note that 'salient' means within our network 'with well defined contrasted edges', but saliency is a more generic concept of local differences, for *e.g*. in intensity, color, orientations as in Itti *et al.*'s model[11]. We could use other types of cells in the first layer to detect other types of saliency, and provided we apply the same intensity-latency conversion, STDP would still focus on the most salient regions. Saliency is known to drive attention (see [18] for a review). Our model predicts

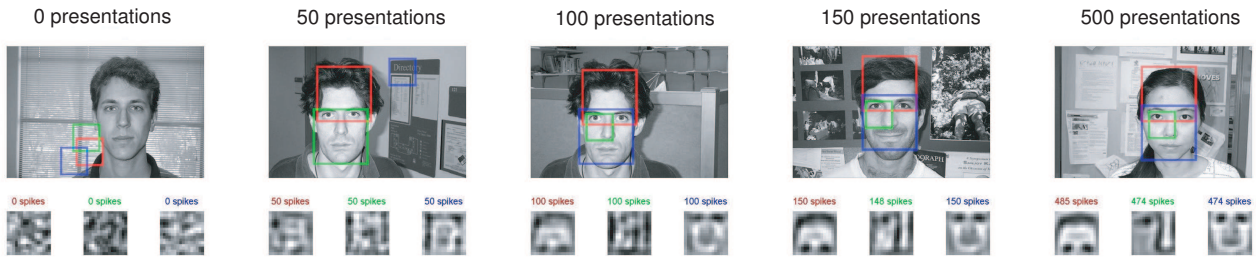


Figure 1. Preferred stimulus reconstructions after 0, 50, 150, and 500 presentations. At the top of each frame the input image is shown, with red, green or blue squares indicating the receptive fields of the cells that fired (if any). At the bottom of the screen we reconstructed the preferred stimuli of the three cells. Above each reconstruction the number of postsynaptic spikes emitted is shown with the corresponding color. See also the video available at http://cerco.ups-tlse.fr/∼masqueli/

that it also drives the learning.

5 Conclusions

We do not claim that our model is realistic: the real ventral stream involves many more layers, and complexity increases more slowly than suggested here. But we think it captures two key mechanisms used by the visual system for rapid object recognition. The first one is the importance of the first spikes to rapidly encode the most important information about a visual stimulus. Given the number of stages involved in high level recognition and the short latencies of selective responses recorded in monkeys' IT[10], the time window available for each neuron to perform its computation is probably around 10-20 ms[17] and will rarely contain more than one or two spikes. The only thing that probably matters for a neuron is whether or not a presynaptic spike is early enough to fall in the critical time window. Later spikes are probably not needed for ultra-rapid categorization. Thus we propose that most of the information must be encoded in the split between the earliest spikes and the others. Clearly such mechanism requires precise stimulus-locked spike times. Recent recordings in monkeys show that IT neurons' responses in terms of spike count close to stimulus onset (100-150 ms time bin) seem to be too reliable to be fit by a typical Poisson firing rate model[2]. There is also experimental evidence for precise spike time responses in V1, and in many other neuronal systems (see[20] for a review).

Very interestingly STDP provides an efficient way to develop selectivity to first spike patterns, as shown in this work. It could explain how a neuron learns to decode the first information available at its afferents' level, to produce fast and selective responses. We believe that STDP is extensively used in the visual system, and constitutes the second key mechanism captured by our model.

6 Acknowledgements

This research was supported by the CNRS, the ACI Neurosciences Computationnelles et Intégratives, and SpikeNet Technology SARL. We also thank Thomas Serre and Rufin VanRullen for reading an earlier version of the manuscript and making comments.

References

[1] D. G. Albrecht, W. S. Geisler, R. A. Frazor, and A. M. Crane. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *J Neurophysiol*, 88(2):888–913., 2002.

[2] A. Amarasingham, T.L. Chen, S. Geman, M.T. Harrison, and D.L. Sheinberg. Spike count reliability and the poisson hypothesis. *J Neurosci*, 26(3):801–809, 2006.

[3] G.Q. Bi and M.M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci*, 18(24):10464–10472, 1998.

[4] S. Celebrini, S. Thorpe, Y. Trotter, and M. Imbert. Dynamics of orientation coding in area v 1 of the awake primate. *Vis Neurosci*, 10(5):811–825, 1993.

[5] R.L. DeValois, D.G. Albrecht, and L.G. Thorell. Spatial frequency selectivity of cells in the macaque visual cortex. *Vision Research*, 22:545–559, 1982.

[6] D.E. Feldman. Timing -based ltp and ltd at vertical inputs to layer ii /iii pyramidal cells in rat barrel cortex. *Neuron*, 27(1):45–56, 2000.

[7] T.J. Gawne, T.W. Kjaer, and B.J. Richmond. Latency : another potential code for feature binding in striate cortex. *J Neurophysiol*, 76(2):1356–1360, 1996.

[8] W. Gerstner and W.M. Kistler. Learning to be fast. In *Spiking neuron models*, pages 421–432. Cambridge University Press, 2002.

[9] R. Guyonneau, R. VanRullen, and S.J. Thorpe. Neurons tune to the earliest spikes through stdp. *Neural Comput*, 17(4):859–879, 2005.

[10] C.P. Hung, G. Kreiman, T. Poggio, and J.J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[11] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[12] H. Markram, J. Lubke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps . *Science*, 275(5297):213–215, 1997.

[13] E.T. Rolls and G. Deco. *Computational neuroscience of vision*. Oxford University Press, 2002.

[14] S. Song, K.D. Miller, and L.F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci*, 3(9):919–926, 2000.

[15] S.J. Thorpe. Spike arrival times: A highly efficient coding scheme for neural networks. In R. Eckmiller, G. Hartmann, and G. Hauske, editors, *Parallel processing in neural systems and computers*, pages 91–94. Elsevier, 1990.

[16] S.J. Thorpe, S. Crouzet, H. Kirchner, and M. Fabre-Thorpe. Ultra-rapid face detection in natural images: implications for computation in the visual system. *Neuro Comp 06*, 2006.

[17] S.J. Thorpe and M. Imbert. Biological constraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Souli, and L. Steels, editors, *Connectionism in perspective*, pages 63–92. Amsterdam: Elsevier, 1989.

[18] S. Treue. Abstract visual attention: the where, what, how and why of saliency. *Curr Opin Neurobiol*, 13(4):428–432, 2003.

[19] M.C.W VanRossum, G.Q. Bi, and G.G. Turrigiano. Stable hebbian learning from spike timing-dependent plasticity. *The Journal of Neuroscience*, 20(23):8812–8821, 2000.

[20] R. VanRullen, R. Guyonneau, and S.J. Thorpe. Spike times make sense. *Trends Neurosci*, 28(1):1–4, 2005.

[21] R. VanRullen and S.J. Thorpe. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput*, 13(6):1255–1283, 2001.

[22] L.I. Zhang, H.W. Tao, C.E. Holt, W.A. Harris, and M. Poo. A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395(6697):37–44, 1998.

B.2.2 Poster

Face feature learning with Spike Timing Dependent Plasticity

Timothée Masquelier & Simon J. Thorpe

Centre de Recherche Cerveau et Cognition (CerCo), Toulouse France

SpikeNet Technology SARL, Labège, France

1. Introduction –

Temporal constraints imposed by both psychophysical and electrophysiological data pose a major challenge to models of face recognition in cortex. Here we explored the capacity of two simple mechanisms. First, when stimulated with a flashed visual stimulus, the neurons in the first layer of the system fire asynchronously, with the most strongly activated neurons firing first. Second, neurons in the second layer implement Spike-Time Dependent Plasticity (STDP).

2. Model –

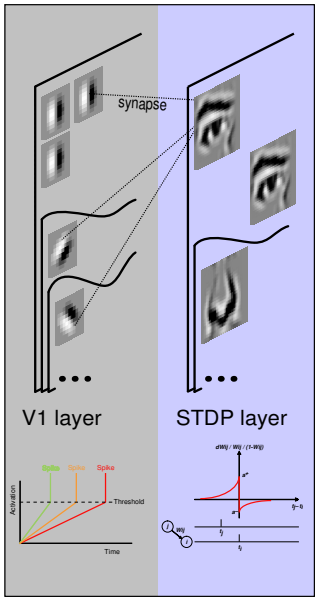
we used a simple 2-layer convolutional feedforward network.

V1 layer

- Oriented edge detectors organized in retinotopic maps
- Intensity-latency conversion. In accordance with experimental data[1-3]. Has been shown to efficiently encode image information[8].
- 1 Winner-Take-All (only the best matching orientation is propagated)

STDP layer

- Spike Timing Dependent Plasticity STDP is a learning rule observed experimentally:
  - Long Term Potentiation (LTP) when a presynaptic spike precedes a postsynaptic one
  - Long Term Depression (LTD) when presynaptic spike follows a postsynaptic one
- STDP is In agreement with Hebb's postulate postulate: presynaptic neurons that fired slightly before the postsynaptic neuron are those which 'took part in firing it'.
- When applied on repetitive similar inputs STDP is known to have the effect of concentrating high synaptic weights on afferents that systematically fire early[5,7].

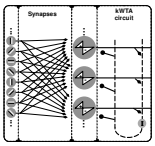


3. Experimental set-up –

we used a Caltech face dataset (available at www.vision.caltech.edu). Faces are seen under various lighting conditions and with very varied backgrounds. Images are presented sequentially and the resulting spike waves are propagated until the STDP layer, where we put 3 prototype neurons. Theses prototype neurons have restricted receptive fields, and are duplicated in retinotopic maps (weight sharing).

We process various scaled versions of the input image (with the same filter size). This results in cells with various receptive field sizes.

Winner-Take-All mechanisms ensures that:



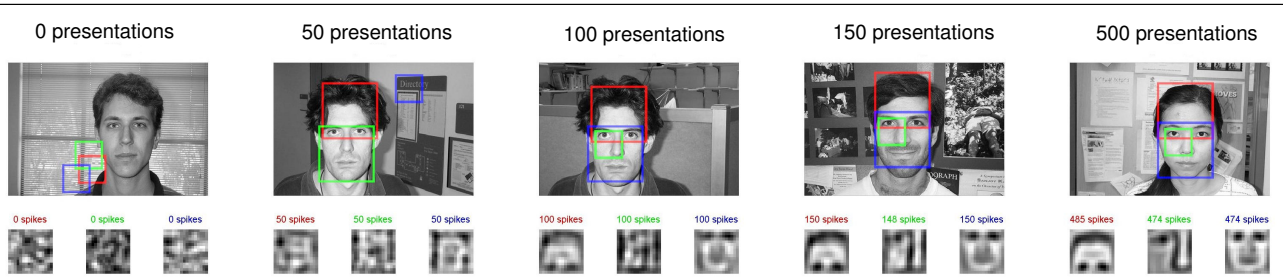
- For each prototype only one duplicated version fires (and trigger the STDP rule). The change in weights is then duplicated at all positions and scales. This allows the system to learn patterns despite of changes in position and size in the training examples.
- At most one cell fires in a given region and scale. This competition prevents all the cells from learning the same pattern.

4. Results –

starting from random preferred stimuli, cells detect statistical regularities among the input spike trains after about one hundred discharges, and progressively develop selectivity to those patterns (see figure below). A few hundred more discharges are needed to reach a stable state, with synapses either maximally potentiated or maximally depressed.

During the learning process:

- synapses compete for control of the timing of postsynaptic spikes, and the winning synapses are those through which the earliest spikes arrive (on average)[5,7]
- the postsynaptic spike latency decreases[4,5,7]
- the algorithm naturally favors salient regions, simply because they correspond to the earliest spikes, with the result that neurons whose receptive fields cover salient regions are likely to reach their threshold (and trigger the STDP rule) before neurons 'looking' at other regions.
- thanks to the competition the cell population self-organizes each neuron learning a distinct pattern, so as to cover the whole variability of the inputs.



Preferred stimulus reconstructions after 0, 50, 150, and 500 presentations. At the top of each frame the input image is shown, with red, green or blue squares indicating the receptive fields of the cells that fired (if any). At the bottom of the screen we reconstructed the preferred stimuli of the three cells. Above each reconstruction the number of postsynaptic spikes emitted is shown with the corresponding color. See also the video available at http://cerco.ups-tlse.fr/~masqueli/

5. Conclusion –

by quickly detecting statistical regularities among the spike trains and creating one unit for each distinct pattern (vector quantization), STDP removes redundancy in the input and performs dimension reduction. It thereby facilitates subsequent processing, for e.g. face recognition. The mechanism is robust enough to deal with natural images and yet leads to highly selective face feature detectors.

Interestingly STDP naturally leads to an 'early spike versus later spike' neural code, which has the advantage of rapidly transmitting information: each neuron produces a reliable response when only a few percent of its afferent have fired. The use of such a mechanism at each stage of the ventral stream could explain the fast and selective responses observed in IT[6].

References

[1] D. G. Albrecht,W. S. Geisler, R. A. Frazor, and A.M. Crane. J Neurophysiol, 2002.

[2] S. Celebrini, S.J. Thorpe, Y. Trotter, and M. Imbert. Vis Neurosci, 1993.

[3] T.J. Gawne, T.W. Kjaer, and B.J. Richmond. J Neurophysiol, 1996.

[4] W. Gerstner and W.M. Kistler. Cambridge University Press, 2002.

[5] R. Guyonneau, R. VanRullen, and S.J. Thorpe. Neural Comput, 2005.

[6] C.P. Hung, G. Kreiman, T. Poggio, and J.J. DiCarlo. Science, 2005.

[7] S. Song, K.D. Miller, and L.F. Abbott. Nat Neurosci, 2000.

[8] R. VanRullen and S.J. Thorpe. Neural Comput, 2001.

B.3 Learning simple and complex cells-like receptive fields from natural images: A plausibility proof

B.3.1 Abstract

Masquelier, T., Serre, T., Thorpe, S., & Poggio, T. (2007). Learning simple and complex cells-like receptive fields from natural images: a plausibility proof [Abstract]. *Journal of Vision*, 7(9):81, 81a, `http://journalofvision.org/7/9/81/`, doi:10.1167/7.9.81.

The ventral stream of the primate's visual system, involved in object recognition, is mostly hierarchically organised. Along the hierarchy (from V1, to V2, V4, and IT) the complexity of the preferred stimulus of the neurons increases, while, at the same time, responses are more and more invariant to shift, scale, and finally viewpoint. Several feedforward networks have been proposed to model this hierarchy by alternating simple cells, which increase selectivity, with complex cells, which increase invariance (Fukushima, 1980; LeCun and Bengio, 1998; Riesenhuber and Poggio, 1999; Serre et al., 2005a). The issue of learning is perhaps the least well understood, and many authors use hard-wired connectivity and/or weight-sharing. Several algorithms have been proposed for complex cell learning based on a trace rule to exploit the temporal continuity of the world (for *e.g.* (Földiák, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Spratling, 2005), but very few can learn from natural cluttered image sequences. Here we propose a new variant of the trace rule that only reinforces the synapses between the most active cells, and therefore can handle cluttered environments. The algorithm has so far been developed and tested through the level of V1-like simple and complex cells: we showed how Gabor-like simple cell selectivity could emerge from competitive hebbian learning, and how the modified trace rule allow the subsequent complex cells to pool over simple cells with the same preferred orientation, but with shifted receptive fields. Development of the V2, V4, and IT layers is ongoing.

B.3.2 Poster

Learning simple and complex cells-like receptive fields from natural images: a plausibility proof

Timothée Masquelier, Thomas Serre, Simon J Thorpe and Tomaso Poggio

Centre de Recherche Cerveau et Cognition (CerCo), Toulouse France
Dept. of Brain and Cognitive Sciences and McGovern Institute for Brain Research, MIT, Cambridge, MA

1. Introduction –

The ventral stream of the primate's visual system, involved in object recognition, is mostly hierarchically organized. Along the hierarchy (from V1, to V2, V4, and IT) the complexity of the preferred stimulus of the neurons increases, while, at the same time, responses are more and more invariant to shift, scale, and finally viewpoint. Several feedforward networks have been proposed to model this hierarchy by alternating simple cells, which increase selectivity, with complex cells, which increase invariance (Fukushima 1980; Le Cun & Bengio 1998; Riesenhuber & Poggio 1999; Serre et al 2007; Masquelier & Thorpe 2007). The issue of learning is perhaps the least well understood, and many authors use hard-wired connectivity and/or weight-sharing. Several algorithms have been proposed for complex cell learning based on a trace rule to exploit the temporal continuity of the world (for e.g., Foldiak 1991; Wallis & Rolls, 1997; Wiskott & Sejnowski, 2002; Einhaüser et al 2002; Spratling 2005), but very few can learn from natural cluttered image sequences.

Here we propose a new variant of the trace rule that, thanks to Winner-Take-All (WTA) competitive mechanisms, only reinforces the synapses between the most active cells, and therefore can handle cluttered environments. We test it on simple cells that learnt their selectivity through a competitive hebbian learning mechanism.

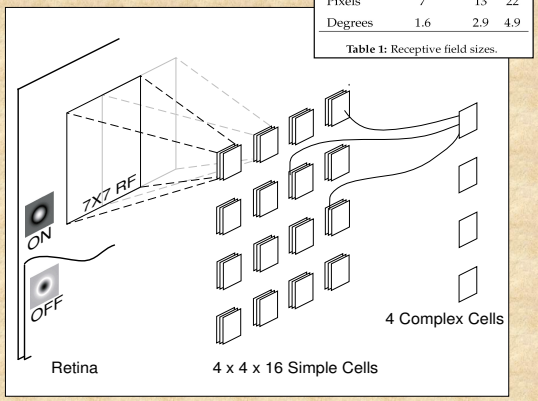
2. Stimuli: the world from a cat's perspective

- Same as in Betsch et al 2004
- CCD cameras attached to cats' heads
- Animals explore several outdoor environments
- Approximate the input which the visual system is naturally exposed to (although eye movements are not taken into account)
- Visual angle: 71 by 53°
- Resolution: 320x240
- A total of 19h of video

3. V1 Model

	ON − OFF	S	C
Pixels	7	13	22
Degrees	1.6	2.9	4.9

Table 1: Receptive field sizes.



4. Retinal ON-OFF cell layer: DoG convolution

- Convolution with 7x7 Difference-of-Gaussian kernel:
- Mimics biological retinal ganglion cells

$$DoG = \frac{1}{2\pi}\left(\frac{1}{\sigma_1}\, e^{-\frac{r^2}{2\sigma_1^2}} - \frac{1}{\sigma_2}\, e^{-\frac{r^2}{2\sigma_2^2}}\right)$$

5. Simple cell layer: competitive hebbian learning

- 16 simple (S) cells in each of the 4x4 cortical columns, starting with random synaptic weights
- Each S cell first computes a normalized dot product between its input x (a 7x7 patch of ON-OFF values) and its synaptic weight vector w

$$y_{\text{raw}} = \frac{\boldsymbol{w}.\boldsymbol{x}}{||\boldsymbol{x}||}$$

- This value is then normalized using a trace (smooth temporal average) of the last dot product values

$$y = \frac{y_{\text{raw}}}{tr(y_{\text{raw}})}$$

- In each cortical column a 1-Winner-Take-All mechanism applies
- The winner will trigger a hebbian rule iff its activity y is above its threshold T:

$$\Delta \boldsymbol{w} = \alpha \cdot y \cdot (\boldsymbol{x} - \boldsymbol{w})$$

- Its threshold is then set to its activity y
- At each time step all the thresholds decrease by 3e-3%

6. Complex cell layer: pool together consecutive S winners

- 4 complex (C) cells receive inputs from the 4x4x16 S cells, through synapses with weights w (0≤w≤1), initially random
- Each C cell computes the maximum value of its weighted inputs:

$$y = \max_{j=1\ldots\, n_{S1}} w_j \cdot x_j$$

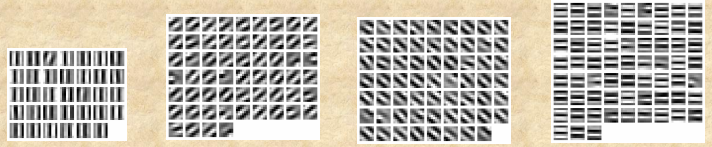
- WTA mechanisms select the C winner at time step t-Δt, J(t-Δt), and the S winner at time t, I(t). The synapse between them is reinforced. All the other synapses of J(t-Δt) are depressed:

$$\Delta w_{ij} = \begin{cases} a^+ \cdot w_{ij} \cdot (1 - w_{ij}) & \text{if } i = I_t \text{ and } j = J_{t-\Delta t} \\ a^- \cdot w_{ij} \cdot (1 - w_{ij}) & \text{otherwise} \end{cases}$$

- Synaptic weights for non-winning C cells are unchanged

7. Results: orientation selectivity, pools with same preferred orientation

- **Simple cells** learn spatial correlations within a frame. Gabor-like orientation selectivity emerges.
- **Complex cells** learn temporal correlations between frames. After convergence weights are binaries (i.e. presence or absence of connection). Here we represent the pools of simple cells each complex cell has formed connection with:



Each C cell pooled together S cells with the same preferred orientation, but at different locations. By taking the maximum value among the pool C cells have a shift-invariant response.

8. Discussion

- Idea: two consecutive S winners are likely to represent the same orientation, so pool them together. This solves the correspondence problem when multiple objects are present even if a given edge activates two simple cells (unlike Spratling 2005)
- Makes this model the only one to our knowledge that handle natural cluttered images, with the one by Einhäuser et al 2002. However:
  - We end up with binary S-C synaptic weights
  - We learn about 20 times faster
  - Our model is simpler (no preprocessing, no lateral inhibition)
  - We have more realistic RF sizes, in particular the RF of our complex cells are about twice as big as the ones of simple cells, leading to a larger shift-invariance
- Biologically plausible. Future work will implement the model on integrate-and-fire neurons
- Development for higher order neurons is ongoing

References

B.Y. Betsch, W. Einhäuser, K.P. Körding, and P. König. Biol. Cybern., 90(1):41–50, 2004.

W. Einhäuser, C. Kayser, P. König, and K.P. Körding. Europ. J. of Neurosc., 15:475–486, 2002

P. Földiák. Neural Comp., 3:194–200, 1991

K. Fukushima. Biol. Cyb., 36:193–202, 1980

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Proc. of the IEEE, 86(11):2278–2324, 1998

T. Masquelier T and S.J Thorpe PLoS Comput. Biol. 3(2), 2007

M. Riesenhuber and T. Poggio. Nat. Neurosci. 2:1019–1025, 1999

T. Serre, A. Oliva and T. Poggio. PNAS, 104(15), pp. 6424-6429, 2007

M.W. Spratling. IEEE Transactions on PAMI, 27(5):753761, 2005

G. Wallis and E.T. Rolls. Prog Neurobiol, 51(2):167–194, 1997

L. Wiskott and T.J. Sejnowski. Neural Computation, 14(4):715770, 2002

List of Tables

List of Figures

Bibliography

Abeles, M. (1982). Role of the cortical neuron: integrator or coincidence detector? *Isr J Med Sci.*, 18(1):83–92.

Abeles, M. (1991). *Corticonics : neural circuits of the cerebral cortex*. Cambridge University Press, Cambridge ; New York. Jean Bullier + Pascal Girard.

Abeles, M. (2004). Neuroscience. time is precious. *Science*, 304(5670):523–4.

Abeles, M. and Gat, I. (2001). Detecting precise firing sequences in experimental data. *J Neurosci Methods*, 107(1-2):141–54.

Adrian, E. (1928). *The Basis of Sensation*. Christophers.

Albrecht, D. G., Geisler, W. S., Frazor, R. A., and Crane, A. M. (2002). Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *J Neurophysiol*, 88(2):888–913.

Amarasingham, A., Chen, T., Geman, S., Harrison, M., and Sheinberg, D. (2006). Spike count reliability and the poisson hypothesis. *J Neurosci*, 26(3):801–809.

Amit, Y. and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vis. Res.*, 43(19):2073–2088.

Anzai, A., Peng, X., and Essen, D. C. V. (2007). Neurons in monkey visual area v2 encode combinations of orientations. *Nat Neurosci*, 10(10):1313–1321.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems*, 3:213–251.

Atienza, M., Cantero, J. L., and Stickgold, R. (2004). Posttraining sleep enhances automaticity in perceptual discrimination. *J Cogn Neurosci*, 16(1):53–64.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.*, 61:183–193.

Bacon-Macé, N., Kirchner, H., Fabre-Thorpe, M., and Thorpe, S. J. (2007). Effects of task requirements on rapid natural scene processing: from common sensory encoding to distinct decisional mechanisms. *J Exp Psychol Hum Percept Perform*, 33(5):1013–1026.

Bacon-Mace, N., Mace, M. J., Fabre-Thorpe, M., and Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Res*, 45(11):1459–69.

Bair, W. and Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Comput*, 8(6):1185–1202.

Baker, C., Behrmann, M., and Olson, C. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.*, 5:1210–1216.

Barlow, H. (1961). *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pages 217–234. MIT Press, Cambridge, MA, wa rosenblith edition.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–94.

Bartlett, M. S. and Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network*, 9(3):399–417.

Baylis, G., Rolls, E., and Leonard, C. (1985). Selectivity between faces in the responses of a population of neurons in the cortex of superior temporal sulcus of the macaque monkey. *Brain Res.*, 342:91–102.

Becker, S. (1991). Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, 2:17–33.

Becker, S. and Hinton, G. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163.

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res*, 37(23):3327–3338.

Benzaquén, A. S. (2006). *Encounters with Wild Children: Temptation and Disappointment in the Study of Human Nature*. McGill-Queen's University Press.

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):579–602.

Berry, M. J. and Meister, M. (1998). Refractoriness and neural precision. *J Neurosci*, 18(6):2200–2211.

Berzhanskaya, J., Grossberg, S., and Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spat Vis*, 20(4):337–395.

Betsch, B., Einhäuser, W., Körding, K., and König, P. (2004). The world from a cat's perspective - statistics of natural videos. *Biological Cybernetics*, 90(1):41–50.

Bi, G. and Poo, M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci*, 18(24):10464–10472.

Bi, G. and Poo, M. M. (2001). Synaptic modification by correlated activity : Hebb's postulate revisited. *Ann Rev Neurosci*, 24:139–166.

Bichot, N. P., Rossi, A. F., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area v4. *Science*, 308(5721):529–34.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J Neurosci*, 2(1):32–48.

Booth, M. C. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex*, 8(6):510–523.

Borg-Graham, L. and Fregnac, Y. (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, 393:369–373.

Boynton, G. and Hegdé, J. (2004). Visual cortex: The continuing puzzle of area V2. *Curr. Biol.*, 14:523–524.

Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J. M., Diesmann, M., Morrison, A., Goodman, P. H., Harris, F. C., Zirpe, M., Natschläger, T., Pecevski, D., Ermentrout, B., Djurfeldt, M., Lansner, A., Rochel, O., Vieville, T., Muller, E., Davison, A. P., Boustani, S. E., and Destexhe, A. (2007). Simulation of networks of spiking neurons: A review of tools and strategies. *J Comput Neurosci*, 23(3):349–398.

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superiortemporal sulcus of the macaque. *J. Neurophys.*, 46(2):369–84.

Buracas, G. T., Zador, A. M., DeWeese, M. R., and Albright, T. D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, 20(5):959–969.

Callaway, E. (1998). Visual scenes and cortical neurons: What you see is what you get. *Proc. Nat. Acad. Sci. USA*, 95(7):3344–3345.

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

Cassenaer, S. and Laurent, G. (2007). Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature*, 448(7154):709–713.

Celebrini, S., Thorpe, S., Trotter, Y., and Imbert, M. (1993). Dynamics of orientation coding in area v 1 of the awake primate. *Vis Neurosci*, 10(5):811–825.

Censor, N., Karni, A., and Sagi, D. (2006). A link between perceptual learning, adaptation and sleep. *Vision Res*, 46(23):4071–4074.

Chase, S. M. and Young, E. D. (2007). First-spike latency information in single neurons increases when referenced to population onset. *Proc Natl Acad Sci U S A*, 104(12):5175–5180.

Coppola, D., Purves, H., McCoy, A., and Purves, D. (1998). The distribution of oriented contours in the real world. *Proc. Nat. Acad. Sci. USA*, 95(7):4002–4006.

Crist, R. E., Li, W., and Gilbert, C. D. (2001). Learning to see: experience and attention in primary visual cortex. *Nat. Neurosci.*, 4:519–525.

de Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., Koberle, R., and Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808.

DeAngelis, G., Anzai, A., Ohzawa, I., and Freeman, R. (1995). Receptive field structure in the visual cortex: Does selective stimulation induce plasticity? *Proc. Nat. Acad. Sci. USA*, 92:9682–9686.

deCharms, R. C. and Merzenich, M. M. (1996). Primary cortical representation of sounds by the coordination of action-potential timing. *Nature*, 381(6583):610–613.

Deco, G. and Lee, T. S. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing*, 44:775–781.

Deco, G. and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res*, 44(6):621–42.

Deco, G. and Rolls, E. T. (2005). Neurodynamics of biased competition and cooperation for attention: a model with spiking neurons. *J Neurophysiol*, 94(1):295–313.

Deco, G. and Zihl, J. (2001). Top-down selective visual attention: A neurodynamical approach. *Visual Cognition*, 8(1):119–140.

Delorme, A., Perrinet, L., Thorpe, S., and M., S. (2001). Networks of integrate-and-fire neurons using rank order coding B: Spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomputing*, 38-40:539–545.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., and Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res Cogn Brain Res*, 19(2):103–113.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2:515–520.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.*, 3:1–8.

Destexhe, A. (1997). Conductance-based integrate-and-fire models. *Neural Computation*, 9(3):503–514.

DeValois, R., Albrecht, D., and Thorell, L. (1982a). Spatial frequency selectivity of cells in the macaque visual cortex. *Vision Research*, 22:545–559.

DeValois, R., Yund, E., and Hepler, N. (1982b). The orientation and direction selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:531–544.

DeWeese, M. R., Wehr, M., and Zador, A. M. (2003). Binary spiking in auditory cortex. *J Neurosci*, 23(21):7940–7949.

DeYoe, E. A. and Essen, D. C. V. (1988). Concurrent processing streams in monkey visual cortex. *Trends. Neurosci.*, 11(5):219–26.

Dill, M. and Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Percept Psychophys.*, 60(1):65–81.

Dolan, R., Fink, G., Rolls, E., Booth, M., Holmes, A., Frackowiak, R., and Friston, K. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature*, 389(6651):596–599.

Douglas, R. J. and Martin, K. A. (1991). A functional microcircuit for cat visual cortex. *J. Physiol. (Lond).*, 440:735–69.

Douglas, R. J. and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annu Rev Neurosci*, 27:419–51.

Einhäuser, W., Kayser, C., König, P., and Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur J Neurosci*, 15(3):475–486.

Erickson, C. A., Jagadeesh, B., and Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nat. Neurosci.*, 3:1143–1148.

Fabre-Thorpe, M., Delorme, A., Marlot, C., and Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci*, 13(2):171–180.

Fabre-Thorpe, M., Richard, G., and Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, 9(2):303–8.

Fahle, M., Edelman, S., and Poggio, T. (1995). Fast perceptual learning in hyperacuity. *Vision Res*, 35(21):3003–3013.

Feldman, D. (2000). Timing -based LTP and LTD at vertical inputs to layer II /III pyramidal cells in rat barrel cortex. *Neuron*, 27(1):45–56.

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

Fellous, J. M., Tiesinga, P. H., Thomas, P. J., and Sejnowski, T. J. (2004). Discovering spike patterns in neuronal responses. *J Neurosci*, 24(12):2989–3001.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. *CVPR*, 2.

Ferster, D. and Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience*, 23:441–471.

Fiorentini, A. and Berardi, N. (1981). Learning in grating waveform discrimination: specificity for orientation and spatial frequency. *Vision Res*, 21(7):1149–1158.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., and Benson, V. (2007). Rapid detection of person information in a naturalistic scene. *Perception*, under press.

Foffani, G., Tutunculer, B., and Moxon, K. A. (2004). Role of spike timing in the forelimb somatosensory cortex of the rat. *J Neurosci*, 24(33):7266–7271.

Földiák, P. (1990). Forming sparse representations by local anti-hebbian learning. *Biol Cybern*, 64(2):165–170.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Földiák, P. (1998). Learning constancies for object perception. In Walsh, V. and Kulikowski, J. J., editors, *Perceptual Constancy: Why things look as they do*, pages 144–172. Cambridge Univ. Press, Cambridge, UK.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci*, 23(12):5235–5246.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cereb Cortex*, 16(11):1631–1644.

Frégnac, Y., Shulz, D., Thorpe, S., and Bienenstock, E. (1988). A cellular analogue of visual cortical plasticity. *Nature*, 333(6171):367–370.

Frégnac, Y. and Shulz, D. E. (1999). Activity-dependent regulation of receptive field properties of cat area 17 by supervised hebbian learning. *J Neurobiol*, 41(1):69–82.

Fries, P., Neuenschwander, S., Engel, A. K., Goebel, R., and Singer, W. (2001). Rapid feature selective neuronal synchronization through correlated latency shifting. *Nat Neurosci*, 4(2):194–200.

Fries, P., Nikolić, D., and Singer, W. (2007). The gamma cycle. *Trends Neurosci*, 30(7):309–316.

Frostig, R. D., Frostig, Z., and Harper, R. M. (1990). Recurring discharge patterns in multiple spike trains. *Biol Cybern*, 62(6):487–493.

Fukushima, K. (1980). Neocognitron : a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*, 36(4):193–202.

Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W., and van Essen, D. C. (1996). Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. *J. Neurophys.*, 76:2718–2739.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, 2(6):568–73.

Gautrais, J. and Thorpe, S. (1998). Rate coding versus temporal order coding: a theoretical approach. *Biosystems*, 48(1-3):57–65.

Gawne, T., Kjaer, T., and Richmond, B. (1996). Latency : another potential code for feature binding in striate cortex. *J Neurophysiol*, 76(2):1356–1360.

Gawne, T. J. and Martin, J. M. (2002). Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophys.*, 88:1128–1135.

Gerstner, W. (1995). Time structure of the activity in neural network models. *Phys. Rev. E*, 51:738–758.

Gerstner, W., Kempter, R., van Hemmen, J. L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76–81. Simon Thorpe.

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models*. Cambridge Univ. Press.

Ghose, G. M. (2004). Learning in mammalian sensory cortex. *Curr Opin Neurobiol*, 14(4):513–518.

Ghose, G. M., Yang, T., and Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *J. Neurophys.*, 87:1867–1888.

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

Girard, P., Jouffrais, C., and Kirchner, H. (2007). Ultra-rapid categorisation in non-human primates. *Animal Cognition*, submitted.

Gross, C. (1972). *Handbook of Sensory Physiology*, volume III, Part 3B, chapter Visual functions of inferotemporal cortex, pages 451–482. Berlin: Springer-Verlag.

Gross, C. (1998). *Brain, Vision, Memory: tales in the history of neuroscience*. MIT Press.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverbarating neural networks. *Studies in Applied Mathematics*, 52:213–257.

Gütig, R. and Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing-based decisions. *Nat Neurosci*, 9(3):420–428.

Guyonneau, R. (2006). *Codage par latence et STDP: des stratégies temporelles pour expliquer le traitement visuel rapide*. PhD thesis, Université Toulouse III - Paul Sabatier.

Guyonneau, R., Kirchner, H., and Thorpe, S. J. (2006). Animals roll around the clock: the rotation invariance of ultrarapid visual processing. *J Vis*, 6(10):1008–1017.

Guyonneau, R., VanRullen, R., and Thorpe, S. (2004). Temporal codes and sparse representations: a key to understanding rapid processing in the visual system. *J Physiol Paris*, 98(4-6):487–497.

Guyonneau, R., VanRullen, R., and Thorpe, S. (2005). Neurons tune to the earliest spikes through STDP. *Neural Comput*, 17(4):859–879.

Hasselmo, M., Rolls, E., and Baylis, G. (1989). The role of expression and identity in the face selective response of neurons in the temporal visual cortex of the monkey. *Behavioral Brain Research*, 32:203–218.

Haykin, S. (1994). *Neural Networks. A Comprehensive Foundation*. Prentice Hall.

Heil, P. (1997). Auditory cortical onset responses revisited. i. first-spike timing. *J Neurophysiol*, 77(5):2616–2641.

Hietanen, J. K., Perrett, D. I., Oram, M. W., Benson, P. J., and Dittrich, W. H. (1992). The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. *Exp. Brain Res.*, 89:157–171.

Hochstein, S. and Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36:791–804.

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

Hopfield, J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature*, 376(6535):33–36.

Horng, S. H. and Sur, M. (2006). Visual activity and cortical rewiring: activity-dependent plasticity of cortical networks. *Prog Brain Res*, 157:3–11.

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

Hubel, D. and Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol*, 28:229–289.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate visual cortex. *J. Phys.*, 148:574–591.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Phys.*, 195:215–243.

Hubel, D. H. and Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J Physiol*, 206(2):419–436.

Hung, C., Kreiman, G., Poggio, T., and DiCarlo, J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866.

Hyvärinen, A. and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vis. Res.*, 41(18):2413–2423.

Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D., and Yuste, R. (2004). Synfire chains and cortical songs: temporal modules of cortical activity. *Science*, 304(5670):559–564.

Innocenti, G. M. and Price, D. J. (2005). Exuberance in the development of cortical networks. *Nat Rev Neurosci*, 6(12):955–965.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res*, 40(10-12):1489–1506.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Jacob, V., Brasier, D. J., Erchova, I., Feldman, D., and Shulz, D. E. (2007). Spike timing-dependent synaptic depression in the in vivo barrel cortex of the rat. *J Neurosci*, 27(6):1271–1284.

Jagadeesh, B., Chelazzi, L., Mishkin, M., and Desimone, R. (2001). Learning increases stimulus salience in anterior inferior temporal cortex of the macaque. *J. Neurophys.*, 86:290–303.

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron*, 53(6):891–903.

Johansson, R. S. and Birznieks, I. (2004). First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat Neurosci*, 7(2):170–177.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol*, 58(6):1233–1258.

Karni, A. and Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proc. Natl. Acad. Sci. USA*, 88:4966–4970.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., and Sagi, D. (1994). Dependence on rem sleep of overnight improvement of a perceptual skill. *Science*, 265(5172):679–682.

Katz, L. C. and Shatz, C. J. (1996). Synaptic activity and the construction of cortical circuits. *Science*, 274(5290):1133–1138.

Keysers, C., Xiao, D. K., Földiák, P., and Perrett, D. I. (2001). The speed of sight. *J. Cogn. Neurosci.*, 13:90–101.

Kiani, R., Esteky, H., and Tanaka, K. (2005). Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. *J Neurophysiol*, 94(2):1587–96.

Kirchner, H. and Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11):1762–1776.

Knill, D. and Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Knoblich, U., Bouvrie, J., and T., P. (2007). Biophysical models of neural computation: Max and tuning circuits. Technical report, CBCL Paper, MIT.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol*, 71(3):856–867.

Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophys.*, 80:324–330.

König, P. (1995). How precise is neuronal synchronization? *Neural Computation*, 7:469–485.

Konig, P., Engel, A. K., and Singer, W. (1996). Integrator or coincidence detector? the role of the cortical neuron revisited. *Trends Neurosci*, 19(4):130–7.

Körding, K., Kayser, C., Einhäuser, W., and König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophys.*, 91(1):206–212.

Kouh, M. and Poggio, T. (2004). A general mechanism for cortical tuning: Normalization and synapses can create gaussian-like tuning. Technical Report AI Memo 2004-031 / CBCL Memo 245, MIT.

Kourtzi, Z., Betts, L. R., Sarkheil, P., and Welchman, A. E. (2005). Distributed neural plasticity for shape learning in the human visual cortex. *PLoS Biol*, 3(7):e204.

Kourtzi, Z. and DiCarlo, J. J. (2006). Learning and neural plasticity in visual object recognition. *Curr Opin Neurobiol*, 16(2):152–158.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *J Neurophysiol*, 92(5):2704–2713.

Laplace, P. (1825). *Essai Philosophique sur les probabilités*. Paris: Gauthier-Villars.

LeCun, Y. and Bengio, Y. (1998). Convolutional networks for images, speech, and time series. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 255âE“–258. Cambridge, MA: MIT Press.

Lee, H., Simpson, G. V., Logothetis, N. K., and Rainer, G. (2005). Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. *Neuron*, 45(1):147–156.

Liu, R. C., Tzonev, S., Rebrik, S., and Miller, K. D. (2001). Variability and information in a neural code of the cat lateral geniculate nucleus. *J Neurophysiol*, 86(6):2789–2806.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–563.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Ann. Rev. Neurosci.*, 19:577–621.

Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat Neurosci*, 4(11):1131–1138.

Luczak, A., Barthó, P., Marguet, S. L., Buzsáki, G., and Harris, K. D. (2007). Sequential structure of neocortical spontaneous activity in vivo. *Proc Natl Acad Sci U S A*, 104(1):347–352.

Macé, M. (2006). *Représentations visuelles précoces dans la catégorisation rapide de scènes naturelles chez l'homme et le singe*. PhD thesis, Université Toulouse III - Paul Sabatier.

Mainen, Z. F. and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps . *Science*, 275(5297):213–215.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London*, B(207):187–217.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2000). Microsaccadic eye movements and firing of single cells in the striate cortex of macaque monkeys. *Nat Neurosci*, 3(3):251–8.

Masquelier, T., Guyonneau, R., and Thorpe, S. J. (2008). Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains. *PLoS ONE*, 3(1):e1377.

Masquelier, T., Serre, T., Thorpe, S., and Poggio, T. (2007). Learning complex cell invariance from natural videos: a plausibility proof. *Massachusetts Institute of Technology*, CBCL Paper #269 / MIT-CSAIL-TR #2007-060.

Masquelier, T. and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol*, 3(2):e31.

McLean, J. and Palmer, L. A. (1998). Plasticity of neuronal response properties in adult cat striate cortex. *Vis Neurosci*, 15(1):177–196.

McLelland, D. and Paulsen, O. (2007). Cortical songs revisited: a lesson in statistics. *Neuron*, 53(3):319–321.

Mehta, M. R., Lee, A. K., and Wilson, M. A. (2002). Role of experience and oscillations in transforming a rate code into a temporal code. *Nature*, 417(6890):741–746.

Mehta, M. R., Quirk, M. C., and Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields [see comments]. *Neuron*, 25(3):707–15.

Mel, B. W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp.*, 9:777–804.

Meliza, C. D. and Dan, Y. (2006). Receptive-field modification in rat visual cortex induced by paired visual stimulation and single-cell spiking. *Neuron*, 49(2):183–189.

Mishkin, M., Ungerleider, L., and Macko, K. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.*

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335:817–820.

Mokeichev, A., Okun, M., Barak, O., Katz, Y., Ben-Shahar, O., and Lampl, I. (2007). Stochastic emergence of repeating cortical motifs in spontaneous membrane potential fluctuations in vivo. *Neuron*, 53(3):413–425.

Movshon, J. A. (2000). Reliability of neuronal responses. *Neuron*, 27(3):412–4.

Mu, Y. and Poo, M. M. (2006). Spike timing-dependent LTP/LTD mediates visual experience-dependent plasticity in a developing retinotectal system. *Neuron*, 50(1):115–25.

Mutch, J. and Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. *cvpr*, 1:11–18.

Nazir, T. A. and O'Regan, J. K. (1990). Some results on translation invariance in the human visual system. *Spat Vis*, 5(2):81–100.

Ofan, R. H. and Zohary, E. (2007). Visual cortex activation in bilingual blind individuals during use of native and second language. *Cereb Cortex*, 17(6):1249–1259.

O'Keefe, J. and Recce, M. L. (1993). Phase relationship between hippocampal place units and the eeg theta rhythm. *Hippocampus*, 3(3):317–330.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., and Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *J Neurosci*, 26(50):13025–13036.

Oram, M. and Földiák, P. (1996). Learning generalisation and localisation: competition for stimulus type and receptive field. *Neurocomputing*, 11(2-4):297–321.

Oram, M. and Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *J Neurophysiol*, 68(1):70–84.

Oram, M. and Perrett, D. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7(6/7):945–972.

Pasupathy, A. and Connor, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophys.*, 82:2490–2502.

Pasupathy, A. and Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophys.*, 86(5):2505–2519.

Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.*, 5(12):1332–1338.

Pasupathy, A. and Miller, E. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 24:873–876.

Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., and Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science*, 297(5580):359–365.

Perrett, D., Hietanen, J., Oram, M., and Benson, P. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. Roy. Soc. B*, 335:23–30.

Perrett, D., Mistlin, A., and Chitty, A. (1987). Visual neurones responsive to faces. *Trends in Neuroscience*, 10:358–364.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Img. Vis. Comput.*, 11:317–333.

Perrett, D., Oram, M., Harries, M., Bevan, R., Hietanen, J., Benson, P., and Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Exp. Brain Res.*, 86:159–173.

Perrett, D., Rolls, E., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res*, 47(3):329–342.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., and Jeeves, M. (1984). Neurones responsive to faces in the temporal cortex: Studies of functional organisation, sensitivity to identity, and relation to perception. *Human Neurobiology*, 3:197–208.

Perrinet, L., Delorme, A., M., S., and Thorpe, S. (2001). Networks of integrate-and-fire neuron using rank order coding A: How to implement spike time dependent hebbian plasticity. *Neurocomputing*, 38-40:817–822.

Perrinet, L., M., S., and Thorpe, S. (2004a). Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. *Neurocomputing*, 57:125–134.

Perrinet, L., Samuelides, M., and Thorpe, S. (2004b). Coding static natural images using spiking event times: do neurons cooperate? *IEEE Trans Neural Netw*, 15(5):1164–1175.

Petersen, R. S., Panzeri, S., and Diamond, M. E. (2001). Population coding of stimulus location in rat somatosensory cortex. *Neuron*, 32(3):503–514.

Pfister, J. and Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *The Journal of Neuroscience*, 26(38):9673âE“9682.

Poggio, T. and Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, 431(7010):768–774.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proc. IEEE*, 78(9).

Poggio, T. and Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS)*, 50(5).

Prut, Y., Vaadia, E., Bergman, H., Haalman, I., Slovin, H., and Abeles, M. (1998). Spatiotemporal structure of cortical activity: properties and behavioral relevance. *J Neurophysiol*, 79(6):2857–74.

Rainer, G., Lee, H., and Logothetis, N. K. (2004). The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biol*, 2(2):E44.

Rainer, G. and Miller, E. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, 27:8–10.

Ramachandran, V. S. (1976). Learning-like phenomena in stereopsis. *Nature*, 262(5567):382–384.

Rehn, M. and Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J Comput Neurosci*, 22(2):135–146.

Reichardt, W., Poggio, T., and Hausen, K. (1983). Figure-ground discrimination by relative movement in the visual system of the fly – II: Towards the neural circuitry. *Biol. Cyb.*, 46:1–30.

Reinagel, P. and Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *J Neurosci*, 20(14):5392–5400.

Reinagel, P. and Reid, R. C. (2002). Precise firing events are conserved across neurons. *J Neurosci*, 22(16):6837–6841.

Richards, W., Feldman, J., and Jepson, A. (1992). From features to perceptual categories. In *Proc. British Machine Vision Conference*, pages 99–108.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2(11):1019–1025.

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

Rolls, E. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Res.*, 66(1-2):177–185.

Rolls, E. (2004). *Invariant object and face recognition. In The Visual Neurosciences*, pages 1165–1178. MIT Press, Cambridge, MA.

Rolls, E. and Deco, G. (2002). *Computational neuroscience of vision.* Oxford University Press.

Rolls, E. and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput*, 12(11):2547–2572.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum Neurobiol*, 3(4):209–22.

Rolls, E. T., Franco, L., Aggelopoulos, N. C., and Jerez, J. M. (2006). Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. *Vision Res*, 46(25):4193–4205.

Rolls, E. T. and Stringer, S. M. (2006). Invariant visual object recognition: a model, with lighting invariance. *J Physiol Paris*, 100(1-3):43–62.

Rousselet, G., Thorpe, S., and Fabre-Thorpe, M. (2003a). Taking the max from neuronal responses. *Trends Cogn Sci*, 7(3):99–102.

Rousselet, G. A., Fabre-Thorpe, M., and Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci*, 5(7):629–30.

Rousselet, G. A., Macé, M. J.-M., and Fabre-Thorpe, M. (2003b). Is it an animal? is it a human face? fast processing in upright and inverted natural scenes. *J Vis*, 3(6):440–455.

Ruderman, D. (1994). The statistics of natural images. *Network : Computation in Neural Systems*, 5:598–605.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Exp Brain Res*, 77(1):23–30.

Schiller, P. H., Finlay, B. L., and Volman, S. F. (1976a). Quantitative studies of single-cell properties in monkey striate cortex I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–1319.

Schiller, P. H., Finlay, B. L., and Volman, S. F. (1976b). Quantitative studies of single-cell properties in monkey striate cortex II. Orientation specificity and ocular dominance. *J. Neurophysiol.*, 39(6):1334–51.

Schiller, P. H., Finlay, B. L., and Volman, S. F. (1976c). Quantitative studies of single-cell properties in monkey striate cortex III. Spatial frequency. *J. Neurophysiol.*, 39(6):1334–1351.

Schoups, A., Vogels, R., Qian, N., and Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412:549–553.

Schuett, S., Bonhoeffer, T., and Hubener, M. (2001). Pairing-induced changes of orientation maps in cat visual cortex. *Neuron*, 32:325–337.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005a). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *Massachusetts Institute of Technology*, CBCL Paper #259/AI Memo #2005-036.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Nat. Acad. Sci. USA*, 104(15).

Serre, T. and Riesenhuber, M. (2004). Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. AI Memo 2004-017 / CBCL Memo 239, MIT, Cambridge, MA.

Serre, T., Wolf, L., and Poggio, T. (2005b). Object recognition with features inspired by visual cortex. *CVPR*, 2:994–1000.

Shiu, L. P. and Pashler, H. (1992). Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Percept Psychophys*, 52(5):582–588.

Sigala, N. and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415(6869):318–20.

Sillito, A. (1984). *Functional properties of cortical cells*, volume 2, chapter Functional considerations of the operation of GABAergic inhibitory processes in the visual cortex, pages 91–117. New York: Plenum Press.

Simoncelli, E. and Olshausen, B. (2001). Natural image statistics and neural representation. *Ann. Rev. Neurosci.*, 24:1193–1216.

Singer, W., Tretter, F., and Yinon, U. (1982). Evidence for long-term functional plasticity in the visual cortex of adult cats. *J. Neurophys.*, 324:239–248.

Sinha, P. and Poggio, T. (1996). The role of learning in 3-D form perception. *Nature*, 384:460–463.

Smith, E. and Lewicki, M. (2006). Efficient auditory coding. *Nature*.

Song, S., Miller, K., and Abbott, L. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci*, 3(9):919–926.

Souza, W. C. D., Eifuku, S., Tamura, R., Nishijo, H., and Ono, T. (2005). Differential characteristics of face neuron responses within the anterior superior temporal sulcus of macaques. *J Neurophysiol*, 94(2):1252–1266.

Spratling, M. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 27(5).

Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: an objective for spike-timing-dependent plasticity? *PLoS Comput Biol*, 3(6):e112.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., and Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. *J Cogn Neurosci*, 12(2):246–254.

Stone, J. and Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network*, 6(3):1–8.

Stringer, S. and Rolls, E. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw*, 13(3):305–315.

Stryker, M. P. (1991). Temporal associations. *Nature*, 354:108–109.

Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.*, 88:135–170.

Swadlow, H. A. and Gusev, A. G. (2002). Receptive-field construction in cortical inhibitory interneurons. *Nat Neurosci*, 5(5):403–404.

Szabo, M., Stetter, M., Deco, G., Fusi, S., Giudice, P. D., and Mattia, M. (2006). Learning to attend: Modeling the shaping of selectivity in inferotemporal cortex in a categorization task. *Biol Cybern*, 94(5):351–65.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci.*, 19:109–139.

Thiele, A., Henning, P., Kubischik, M., and Hoffmann, K.-P. (2002). Neural mechanisms of saccadic suppression. *Science*, 295(5564):2460–2462.

Thorpe, S. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In Eckmiller, R., Hartmann, G., and Hauske, G., editors, *Parallel processing in neural systems and computers*, pages 91–94. Elsevier.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–2. PDF + Simon Thorpe + Rufin.

Thorpe, S. and Imbert, M. (1989). Biological constraints on connectionist modelling. In *Connectionism in perspective*, pages 63–92. Amsterdam: Elsevier.

Thorpe, S. J. and Fabre-Thorpe, M. (2001). Neuroscience. seeking categories in the brain. *Science*, 291(5502):260–263.

Torre, V. and Poggio, T. (1978). A synaptic mechanism possibly underlying directional selectivity motion. *Proc. of the Royal Society London B*, 202:409–416.

Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2005). Generalized bienenstock-cooper-munro rule for spiking neurons that maximizes information transmission. *Proc Natl Acad Sci U S A*, 102(14):5239–5244.

Treue, S. (2003). Abstract visual attention: the where, what, how and why of saliency. *Curr Opin Neurobiol*, 13(4):428–432.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., and Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artificial. Intelligence.*, 78(1-2):507–45.

Uchida, N., Kepecs, A., and Mainen, Z. F. (2006). Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making. *Nat Rev Neurosci*, 7(6):485–491.

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat Neurosci*, 5(7):682–687.

Ungerleider, L. G. and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr. Op. Neurobiol.*, 4:157–165.

Uzzell, V. J. and Chichilnisky, E. J. (2004). Precision of spike trains in primate retinal ganglion cells. *J Neurophysiol*, 92(2):780–789.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc Biol Sci*, 265(1412):2315–2320.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc Biol Sci*, 265(1394):359–366.

VanRossum, M., Bi, G., and Turrigiano, G. (2000). Stable hebbian learning from spike timing-dependent plasticity. *The Journal of Neuroscience*, 20(23):8812–8821.

VanRullen, R., Gautrais, J., Delorme, A., and Thorpe, S. (1998). Face processing using one spike per neurone. *Biosystems*, 48(1-3):229–239.

VanRullen, R., Guyonneau, R., and Thorpe, S. (2005). Spike times make sense. *Trends Neurosci*, 28(1):1–4.

VanRullen, R. and Thorpe, S. (2001). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput*, 13(6):1255–1283.

VanRullen, R. and Thorpe, S. (2002). Surfing a spike wave down the ventral stream. *Vision Res*, 42(23):2593–2615.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 17:264–280.

Vislay-Meltzer, R. L., Kampff, A. R., and Engert, F. (2006). Spatiotemporal specificity of neuronal activity directs the modification of receptive fields in the developing retinotectal system. *Neuron*, 50(1):101–14.

Wachsmuth, E., Oram, M., and Perrett, D. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4:509–522.

Walker, M. P., Stickgold, R., Jolesz, F. A., and Yoo, S.-S. (2005). The functional anatomy of sleep-dependent visual skill learning. *Cereb Cortex*, 15(11):1666–1675.

Wallis, G. (1996). Using spatio-temporal correlations to learn invariant object recognition. *Neural Networks*, 9(9):1513–1519.

Wallis, G. and Bülthoff, H. (2001). Role of temporal association in establishing recognition memory. *Proc. Nat. Acad. Sci. USA*, 98(8):4800–4804.

Wallis, G. and Rolls, E. (1997). Invariant face and object recognition in the visual system. *Prog Neurobiol*, 51(2):167–194.

Wallis, G., Rolls, E., and Földiák, P. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks*, 2:1087–1090.

Wersing, H. and Koerner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Comp.*, 15(7):1559–1588.

Wilson, F., Scalaidhe, S., and Goldman-Rakic, P. (1994). Functional synergism between putative y-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. *Proc. Nat. Acad. Sci. USA*, 91:4009–4013.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput*, 14(4):715–770.

Yamane, S., Kaji, S., and Kawano, K. (1988). What facial features activate face neurons in the inferior temporal cortex of the monkey? *Exp. Brain Res.*, 73:209–214.

Yang, T. and Maunsell, J. H. R. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *J. Neurosci.*, 24:1617–1626.

Yao, H. and Dan, Y. (2001). Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron*, 32:315–323.

Young, J. M., Waleszczyk, W. J., Wang, C., Calford, M. B., Dreher, B., and Obermayer, K. (2007). Cortical reorganization consistent with spike timing – but not correlation-dependent plasticity. *Nat. Neurosc.*, 10(7):887–895.

Yu, A. J., Giese, M. A., and Poggio, T. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Comp.*, 14(12):2857–2881.

Zhang, L., Tao, H., Holt, C., Harris, W., and Poo, M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395(6697):37–44.

Zoccolan, D., Cox, D. D., and DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *J Neurosci*, 25(36):8150–8164.

Learning mechanisms to account for the speed, selectivity and invariance of responses in the visual cortex

In this thesis I propose various activity-driven synaptic plasticity mechanisms that could account for the speed, selectivity and invariance of the neuronal responses in the visual cortex. Their biological plausibility is discussed. I also present the results of a relevant psychophysical experiment demonstrating that familiarity can accelerate visual processing. Beyond these results on the visual system, the studies presented here also credit the hypothesis that the brain uses the spike times to encode, decode, and process information – a theory referred to as 'temporal coding'. In such a framework the Spike Timing Dependent Plasticity may play a key role, by detecting repeating spike patterns and by generating faster and faster responses to those patterns.

Auteur	Timothée Masquelier
Titre	Mécanismes d'apprentissage pour expliquer la vitesse, la sélectivité et l'invariance des réponses dans le cortex visuel
Directeur de thèse	Simon J Thorpe
Lieu et date de soutenance	Le 15 février 2008 à l'Université Paul Sabatier Toulouse III

Résumé

Dans cette thèse je propose plusieurs mécanismes de plasticité synaptique qui pourraient expliquer la rapidité, la sélectivité et l'invariance des réponses neuronales dans le cortex visuel. Leur plausibilité biologique est discutée. J'expose également les résultats d'une expérience de psychophysique pertinente, qui montrent que la familiarité peut accélérer les traitements visuels. Au delà de ces résultats propres au système visuel, les travaux présentés ici créditent l'hypothèse de l'utilisation des dates de spikes pour encoder, décoder, et traiter l'information dans le cerveau – c'est la théorie dite du 'codage temporel'. Dans un tel cadre, la Spike Timing Dependent Plasticity pourrait jouer un rôle clef, en détectant des patterns de spikes répétitifs et en permettant d'y répondre de plus en plus rapidement.

Mots-cles

vision, reconnaissance d'objets, catégorisation visuelle ultra-rapide, apprentissage, codage temporel, neurones impulsionnels, Spike Timing Dependent Plasticity (STDP)

Discipline administrative

Neurosciences Cognitives

Laboratoire

Centre de Recherche Cerveau et Cognition UMR 5549 (CNRS-Université Paul Sabatier Toulouse III), Faculté de Médecine de Rangueil 31062 Toulouse CEDEX9