Learning complex cell invariance from natural videos: A plausibility proof
Timothee Masquelier, Thomas Serre, Simon Thorpe, and Tomaso Poggio
Learning complex cell invariance from natural videos: A plausibility proof

Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA, 02139

Approved for public release; distribution unlimited
One of the most striking features 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; 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 extra-striate 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. In fact, in most of these models, the connectivity between simple and complex cells is not learned but rather hard-wired. Several algorithms have been proposed for learning invariances at the complex cell level based on a trace rule to exploit the temporal continuity of sequences of natural images (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 at the level of V1-like simple and complex cells: we verified that Gabor-like simple cell selectivity could emerge from competitive Hebbian learning (see also (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006)). In addition, we show how the modified trace rule allows the subsequent complex cells to learn to selectively pool over simple cells with the same preferred orientation but slightly different positions thus increasing their tolerance to the precise position of the stimulus within their receptive fields.
Learning complex cell invariance from natural videos: a plausibility proof

Timothée Masquelier, Thomas Serre, Simon Thorpe and Tomaso Poggio
Abstract

One of the most striking features 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; Berzhanskiya 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; VanRullen et al., 1999; 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 extra-striate 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. In fact, in most of these models, the connectivity between simple and complex cells is not learned but rather hard-wired. Several algorithms have been proposed for learning invariances at the complex cell level based on a trace rule to exploit the temporal continuity of sequences of natural images (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 at the level of V1-like simple and complex cells: we verified that Gabor-like simple cell selectivity could emerge from competitive Hebbian learning (see also (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006)). In addition, we show how the modified trace rule allows the subsequent complex cells to learn to selectively pool over simple cells with the same preferred orientation but slightly different positions thus increasing their tolerance to the precise position of the stimulus within their receptive fields.
1 Introduction

Learning is arguably the key to understanding intelligence (Poggio and Smale, 2003). One of the most striking features of the cortex is its ability to wire itself. 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 and Oram, 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 (RFs) of neurons in the visual cortex.

Here we study biologically plausible mechanisms for the learning of both 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; 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.

We have used a specific implementation of such feedforward hierarchical models (Riesenhuber and Poggio, 1999; Serre et al., 2005b, 2007), which makes predictions about the nature of the computations and the specific wiring of simple and complex units, denoted $S_1$ and $C_1$ units respectively. 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 the Gabor-like selectivity of $S_1$ units could emerge from competitive Hebbian learning (see also (Delorme et al., 2001; Einhäuser et al., 2002; Guyonneau, 2006)). In addition, we proposed a new mechanism, which suggests how the specific pooling from $S_1$ to $C_1$ unit 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.

1.1 Evidence for 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 debate.

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).

At the cellular level, supervised learning procedures to validate Hebb’s covariance hypothesis in vivo 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). Indeed it was 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 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 cells in the visual cortex.
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 (RFs). 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. 1 (reproduced from (Hubel and Wiesel, 1962)), 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 1: The Hubel & Wiesel hierarchical model for building complex cells from simple cells. Reproduced from (Hubel and Wiesel, 1962).

1.2 Simple and complex cell modeling

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 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, 1962) (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:

The simple $S$ units perform a TUNING operation over their afferents to build object-selectivity. 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).

The analog of the TUNING operation in computer vision is the template matching operation between an input image and a stored representation. 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).
The complex C units receive convergent inputs from retinotopically organized S units tuned to the same preferred stimuli but at slightly different positions and scales with 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 a soft-max close to an average may be sufficient for robust and invariant object recognition (Serre and Poggio, 2005) and seems to account for a significant proportion of complex cells (Finn and Ferster, 2007). Fig. 2 (reproduced from (Lampl et al., 2004)) illustrates how a complex cell may combine the response of oriented retinotopically organized subunits (presumably simple cells) at the same preferred orientation with a MAX pooling mechanism.

Computational implementation of the Hubel & Wiesel model: In this work we use static idealized approximation to describe the response of simple and complex units. As described in (Knoblich et al., 2007; Kouh and Poggio, 2007; 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).
We denote by \((x_1, x_2, \ldots, x_n)\) the set of inputs to a unit and \(w = (w_1, \ldots, w_N)\) their respective input strength. For a complex unit, the inputs \(x_j\) are retinotopically organized and selected from an \(m \times m\) grid of afferent units. Of afferents 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:

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

where \(y\) is the output of the unit, \(k \ll 1\) 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 \(p \lesssim qr\), the unit approximates a Gaussian-like TUNING, i.e., its response \(y\) will have a peak around some value proportional to the input vector \(w = (w_1, \ldots, w_N)\). 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, 2007; Serre et al., 2005a) for details).
- When \(p \gtrsim q + 1\) (\(w_j \approx 1\)), 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 (Kouh and Poggio, 2007; Serre et al., 2005a) for details).

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 population of spiking neurons firing probabilistically in proportion to the underlying analog value (Smith and Lewicki, 2006) and on shunting inhibition (Grossberg, 1973). Other possibilities may involve spike timing in individual neurons (Masquelier and Thorpe, 2007) (see (VanRullen et al., 2005) for a recent review). A complete description of the two operations, a summary of the evidence as well as plausible biophysical circuits to implement them can be found in (Knoblich et al., 2007; Serre et al., 2005a).

While there exists at least partial evidence for the existence of both Gaussian TUNING and max-like operations (see earlier), 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.

### 1.3 On learning simple and complex cells

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; Földiáik, 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 (Serre et al., 2005a).
The wiring of the simple units corresponds to learning 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 (von der Malsburg, 1973; 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 Section 4, Eq. 5). 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 4.3). 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, horizontal connections are unlikely to play a critical role during 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). They 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, which could also be implemented in the visual cortex via horizontal connections, has been proposed to reduce the redundancy and increases the sparseness of neuronal responses (Perrinet et al., 2004).

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 with a specific implementation of a model of the ventral stream of the visual cortex (Riesenhuber and Poggio, 1999; Serre et al., 2005a, 2007), formerly known as HMAX.

The wiring of complex units, on the other hand, may reflect learning from visual experience how to associate frequent transformations in time – such as translation and scale – of specific image 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).

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 positions and scales. In this work we focus on translation invariance, but the same mechanism, in principle, could be applied to any transformation (e.g., scale, rotation or viewpoint).

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 coincide 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 work we focus on the learning of simple $S_1$ and complex $C_1$ units (see Fig. 3), which constitutes a direct implementation of the Hubel and Wiesel (1962) 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 results suggests (see later), this constitutes a reasonable hypothesis, which leads to appropriate pooling.

2 Results

We tested the proposed learning mechanisms in a 3 layer feedforward network mimicking the Lateral Geniculate Nucleus (LGN) and V1 (see Fig. 3). Details of the implementation can be found in Section 4.

The stimuli we used were provided by Betsch et al. (2004). The videos were captured by CCD cameras attached to a cat’s head, while the animal was 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.
2.1 Simple cells

After about 9 hours of simulated time $S_1$ units have learned a Gabor-like selectivity (see Fig. 4) 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 particular, 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 $45^\circ$ increment: $0^\circ$, $45^\circ$, $90^\circ$ and $135^\circ$. Interestingly, in an another experiment using $S_1$ receptive fields larger than the $7 \times 7$ receptive field sizes used here, we found instead a continuum of orientations. 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 $7 \times 7$ RF sizes (see Table 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) and (Olshausen and Field, 1996) for a more sophisticated model).
2.2 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 4.4). 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). 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. 6). 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(a), 5(b), 5(c), 5(d) and 6 were all disjoint. Note that superimposing those 5 figures leads to Fig. 4.

We also experimented with other learning rules within the same architecture. We reimplemented Földiák’s original trace rule (Földiák, 1991) (see Section 4, Eq. 12). 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 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 Section 4, Eq. 10). 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 3).

3 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 any given algorithm to be plausible a necessary condition (although not sufficient) is that it can handle natural images, which brings 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 ultimate goal 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 9), 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 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 some of them lead to reasonable rules (e.g., (Olshausen and Field, 1996; Sprekeler et al., 2007)) and connections can be drawn between the two classes of approaches. For example, Spreker 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; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2007; Masquelier and Thorpe, 2007), which typically learn the tuning of units at one location and simply ‘replicate’ the tuning of 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.

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 a plausibility proof that invariances could be learned using a simple trace-rule, even in natural cluttered environment.

4 Methods

4.1 Stimuli: the world from a cat’s perspective

The videos used were taken from (Betsch et al., 2004) (see Fig. 7). 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 se-
quences 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 1.

**Table 1:** Receptive field sizes in pixels, and in degree of visual angle.

<table>
<thead>
<tr>
<th></th>
<th>ON−OFF</th>
<th>$S_1$</th>
<th>$C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixels</td>
<td>7</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>Degrees</td>
<td>1.6</td>
<td>2.9</td>
<td>4.9</td>
</tr>
</tbody>
</table>

4.2 LGN ON- and OFF-center unit layer

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

$$ DoG = \frac{1}{2\pi} \left( \frac{1}{\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{-\frac{x^2}{2\sigma_2^2}} \right) $$

(1)

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.

4.3 $S_1$ layer: competitive Hebbian learning

Model $S_1$ units are organized on a $4 \times 4$ grid of cortical columns. Each column contains 16 $S_1$ units (see Fig. 3). The distance between columns was set to 3 pixels (i.e., about half a degree of visual angle). Each $S_1$ unit received its inputs from a $7 \times 7$ 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):

$$ y_{raw} = \frac{\sum_{j=1}^{n} w_j \cdot x_j^p}{k + (\sum_{j=1}^{n} x_j^r)^r} $$

(2)

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:

$$ y_{raw} = \frac{w \cdot x}{||x||} $$

(3)

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 Eq. 3.

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’:

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

(4)

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).

The initial $w$ weights of all the $S_1$ units were initialized at random (sampled from a uniform distribution on the [0,1] interval). In each cortical hypercolumn 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:

$$ \Delta w = \alpha \cdot y \cdot (x - w) $$

(5)

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:

$$ T = y $$

(6)

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

$$ T = (1 - \eta) \cdot T $$

(7)

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:

$$tr(y_{raw}) = \frac{y_{raw}}{\nu} + (1 - \frac{1}{\nu}) \cdot tr(y_{raw})$$  (8)

We used $\eta = 2^{-15}$ and $\nu = 100$. It was found useful to geometrically increase the learning rate $\alpha$ 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.

### 4.4 C1 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 0.75).

Each $C_1$ cell’s activity is computed using Eq. 2, 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:

$$\Delta w_{iJ_{t-\Delta t}} = \begin{cases} a^+ \cdot w_{iJ_{t-\Delta t}} \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{if } i = I_t \\ a^- \cdot w_{iJ_{t-\Delta t}} \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{otherwise.} \end{cases}$$  (9)

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 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 iterations, 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.

### 4.5 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, Einhäuser et al. (2002) use a different weight update rule:

$$\Delta w_{iJ_t} = \begin{cases} \alpha \cdot (1 - w_{iJ_t}) & \text{if } i = I_t \\ -\alpha \cdot w_{iJ_t} & \text{otherwise.} \end{cases}$$  (10)

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 the rule by Einhäuser et al. (2002):

$$\Delta w_{iJ_{t-\Delta t}} = \begin{cases} \alpha \cdot (1 - w_{iJ_{t-\Delta t}}) & \text{if } i = I_t \\ -\alpha \cdot w_{iJ_{t-\Delta t}} & \text{otherwise.} \end{cases}$$  (11)

This suggests that the problem comes from the weight update rule that was used in (Einhäuser et al., 2002), and not from the type of correlation involved.

- In the model by (Einhäuser et al., 2002) the activity of the units is normalized by the “trace” (see earlier) also at the complex cell level.

- The model by (Einhäuser et al., 2002) is similar to an energy-type model of complex cells such that simple and complex cells have identical receptive field sizes. In particular, the model does not account for the increase in RF sizes between $S_1$ and $C_1$ units (typically doubling (Hubel and Wiesel, 1962).

### 4.6 Földiák’s original trace rule

We also reimplemented Földiák’s original trace rule (Földiák, 1991) (see Section 2), which is given by:

$$\Delta w = \alpha \cdot tr(y) \cdot (x - w).$$  (12)

### Notes

1 As suggested by several authors (Földiák and Young, 1995; Perrett et al., 1998; Shadlen and Newsome, 1998; Keysers et al., 2001; Serre et al., 2005a), because of the strong temporal constrains imposed on the cortical circuits (i.e., computations at each stage have to be performed within very small temporal windows of $10 - 30 \text{ ms}$ (Thorpe and Imbert, 1989; Thorpe and Fabre-Thorpe, 2001; Keysers...
et al., 2001; Rolls, 2004) under which single neurons can transmit only very few spikes), the basic units of processing are likely to be modules of hundreds of neurons with similar selectivities rather than individual cells. Such computational modules could be implemented by cortical columns (Mountcastle, 1957, 1997). To paraphrase Mountcastle (Mountcastle, 1997): “the effective unit of operation in such a distributed system is not the single neuron and its axon, but groups of cells with similar functional properties and anatomical connections”.

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.

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).

Acknowledgments

We would like to thank W. Einhäüser for sharing with us the catcam video stimuli used in this work.

References


