ELSEVIER

Neurocomputing 26–27 (1999) 911–918

NEUROCOMPUTING

www.elsevier.com/locate/neucom

# Spatial attention in asynchronous neural networks

Rufin VanRullen*, Simon J. Thorpe

*Centre de Recherche Cerveau et Cognition, Faculté de Médecine de Rangueil, 133, Route de Narbonne, 31062, Toulouse Cedex, France*

Accepted 18 December 1998

**Abstract**

We propose a simple mechanism for spatial visual attention that involves selectively lowering the thresholds of neurons with receptive fields in the attended region. Whereas such a mechanism is of no use in classical artificial neural networks, where all activities for each position in the visual field are computed simultaneously, it can be of great interest in an asynchronous neural network, where the relative order of firing in a population of neurons constitutes the code. Since neurons in the attended region will tend to reach threshold and fire earlier, they will tend to dominate later stages of processing. We illustrate this hypothesis with simulations based on SpikeNET. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Attention; Rank-order coding; Spiking neurons; Threshold lowering

## 1. Introduction

There are numerous different theories and models to explain spatial attention mechanisms in the visual field. But none of them takes account of the asynchrony inherent in real neural networks such as the visual system. Yet it is well known that neurons in a given population fire at different rates, but also at different latencies.

We have already proposed [3,4] that these differences in firing latencies, e.g. the relative order of firings in a population, could be used as a code for transmitting the information from one processing stage to the next. The most strongly activated

* Corresponding author. E-mail: rufin@cerco.ups-tlse.fr.

PII: S0925-2312(98)00136-2

neurons will tend to fire first, with the result that early processing in later stages will be dominated by the shortest latency inputs. Neurons in later processing stages can be made to be sensitive to the order in which their inputs fire by invoking a mechanism which progressively decreases the post-synaptic neuron's sensitivity as more and more inputs arrive [4]. We have demonstrated that it is perfectly conceivable to produce multi-layered feed-forward architectures based on such principles that are capable of performing complex visual processing tasks that include the localization of faces in natural images [6].

Under such conditions, we can make the hypothesis that spatial attention involves selectively lowering the effective threshold of neurons with receptive fields in the attended region. This means that neurons at this location will tend to fire earlier, giving a temporal precedence to the attended stimuli, and allowing them to dominate processing at later stages.

## 2. Why the visual system needs spatial attention

The need for spatial attention, as pointed out by Mozer and Sitton [2] stems from the resource limitations of real visual systems.

Consider a neural network performing object recognition. With one neuron selective to a particular object for each spatial location, such a system does not need any attentional mechanism to perform accurately. For example, we have proposed [6] a model for face detection that does not use attention. The problem arises in real networks such as the human visual system, where the amount of resources, namely the number of neurons, is limited.

Clearly, the human visual system cannot afford one "object detector" for each object and each retinotopic location. It is well known that neurons in the visual system have increasing receptive fields sizes, and many neurons in the latest stages, such as the inferotemporal cortex, have receptive fields covering the entire visual field. They can respond to an object independently of its spatial location. Such a system needs far fewer neurons. But how can it deal with more than one object at the same time? With no attentional mechanism, if you present to that network an image containing many objects to detect, it is impossible to decide which one it will choose. Furthermore, there is a risk that features from different objects will be mixed, causing problems for accurate identification. This is an aspect of the well-known "binding problem" [5]

Suppose now we lower the thresholds for neurons with receptive fields in one part of the visual field. Provided that the neurons have dynamical properties such as those observed in real neurons (integrate-and-fire, spiking neurons …), information concerning the object in this region will tend to propagate more quickly through the network, and so will activate the appropriate output detector before information about the other objects has arrived. The network response will thus correspond to what was in the image at the location of attention.

## 3. Simulations

Here we follow the argumentation of Mozer and Sitton [2] and translate it in the context of asynchronous neural networks. We have shown [6] that this kind of networks are suitable for complex visual tasks like face detection, provided that the amount of neurons used is not a limiting factor. Here we illustrate the problem of resource limitations in the context of object recognition. Finally, we demonstrate that our hypothesis allows the model to overcome these problems.

More precisely, we have built simple object recognition models to explore the possibility that such a threshold-decrease mechanism could underly the effects of spatial attention. These models were implemented with SpikeNET, our large-scale asynchronous neural network simulation software [1].

Units in SpikeNET are simple integrate-and-fire neurons, which basically generate no more than one spike for each image presented to the network. Furthermore, they can be made to be selective to a particular *order* of their afferent spikes, by a mechanism which decreases the neurons sensitivity as more and more inputs arrive, irrespective of their weight. Therefore, the neurons will be best activated when the order of their inputs matches the order of their synaptic weights [4].

Using this particular neural network scheme, we built 2 different models of object recognition, and compared their performance on a very simple categorization task: 9 views of 9 different objects (1 view per object, Fig. 1) were learned, and had to be recognized at any of 4 different locations, corresponding to the left or right and upper or lower hemifields.

The two models shared the same 6-level hierarchical organization. Units in the first level, corresponding to the retina, responded to a positive or negative local contrast (ON- or OFF-Center cells). At that level, the analog intensity of the input contrast was transformed in a firing latency. Units in the second layer were selective for edges of a particular orientation (8 different orientations separated by 45°), like the simple cells of the primary visual cortex V1, whereas the third layer combined these informations in 4 different maps, in which neurons were selective for an orientation irrespective of its polarity. At the next processing stage, basic features like terminations, T- or L-junctions, at 8 possible orientations, were extracted, and then combined using "complex" cells in the 5th layer. Finally, neurons in the last layer were trained to respond specifically to different objects.

The 2 models differ only by the presence or absence of an attentional mechanism.

### 3.1. Limited resources model, without attention

In the first model, as in the visual system, we wanted the sizes of the neurons receptive fields to increase from one processing stage to the next, so that the object detectors receptive fields, as observed in IT, would cover the entire visual field. In this case there is only one neuron per object category in the final layer. That kind of organization required only 72073 neurons, whereas the same hierarchical model without resource limitations would use up to 1146880 neurons. An example of the propagation of an image through that network is shown in Fig. 2.

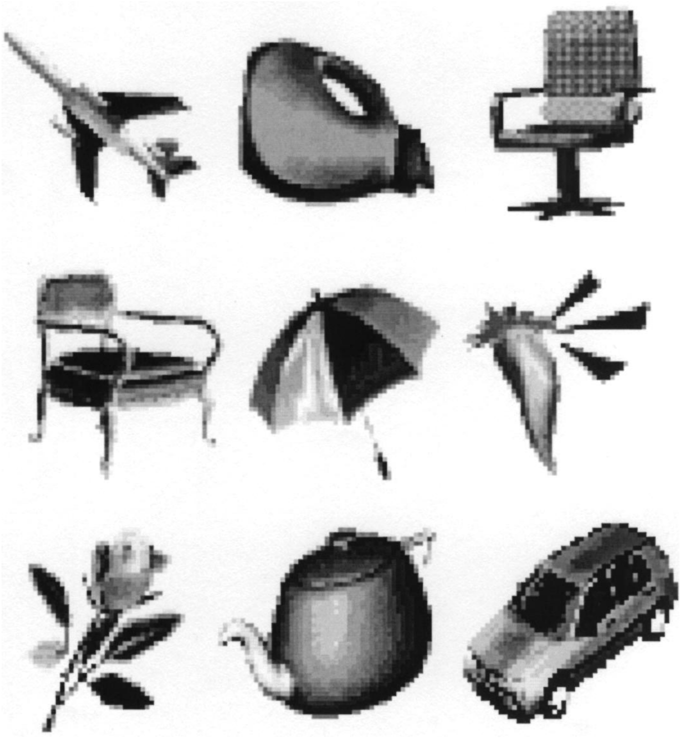


Fig. 1. Objects to be learned and categorized by the different models.

Since we had only one single "object detector" per object, the supervised learning was made as follows: we computed the mean pattern of firing order obtained, in the "complex features layer", for one object presented at each of the four possible locations, and that mean pattern became the order of the weights of the neuron selective for that object.

Furthermore, we introduced a lateral inhibition between the output neurons, so that only the first(s) one(s) to reach their threshold would respond.

Though the computation time was less than 1 s, the performance of that model was really poor. When objects were presented alone, they were always detected, without confusion with other objects. But when the objects were presented by pairs, in only 88% of the images one of the 2 targets was recognized, and in 22% of the trials, a completely different object was detected (see Fig. 2).

As expected, this kind of organization, with increasing receptive fields sizes, is a good way of saving neurons, but makes the model unable to deal with more than one object simultaneously, because features belonging to different objects are likely to be wrongly associated.

Nevertheless, it is well known that this organization scheme is indeed used by the human visual system. We propose that an attentional model in which the thresholds

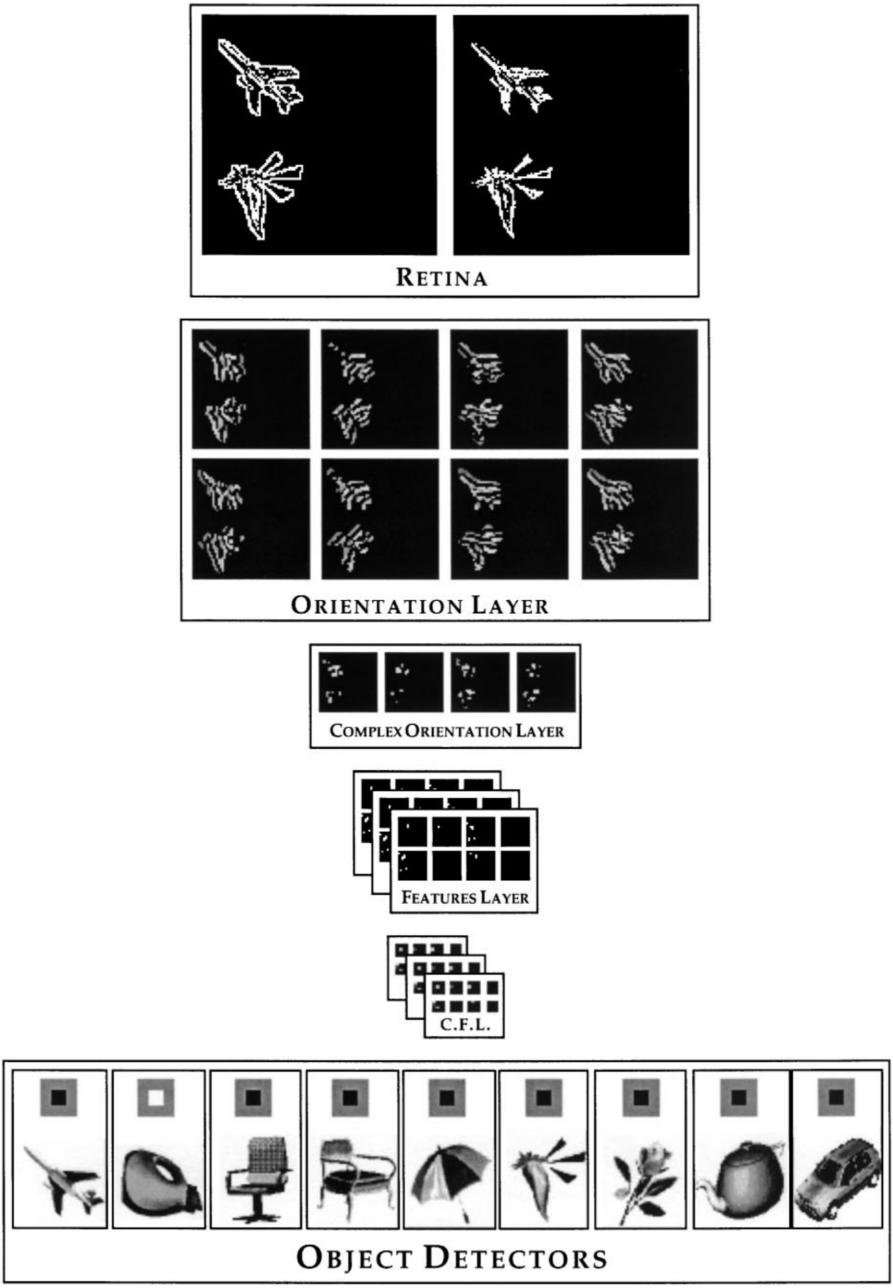


Fig. 2. An example of the propagation of an image in the 1st network. Each pixel in these maps represents a neuron, with white pixels corresponding to activated neurons. The output neurons sizes have been increased. Note that the network outputs a wrong object.

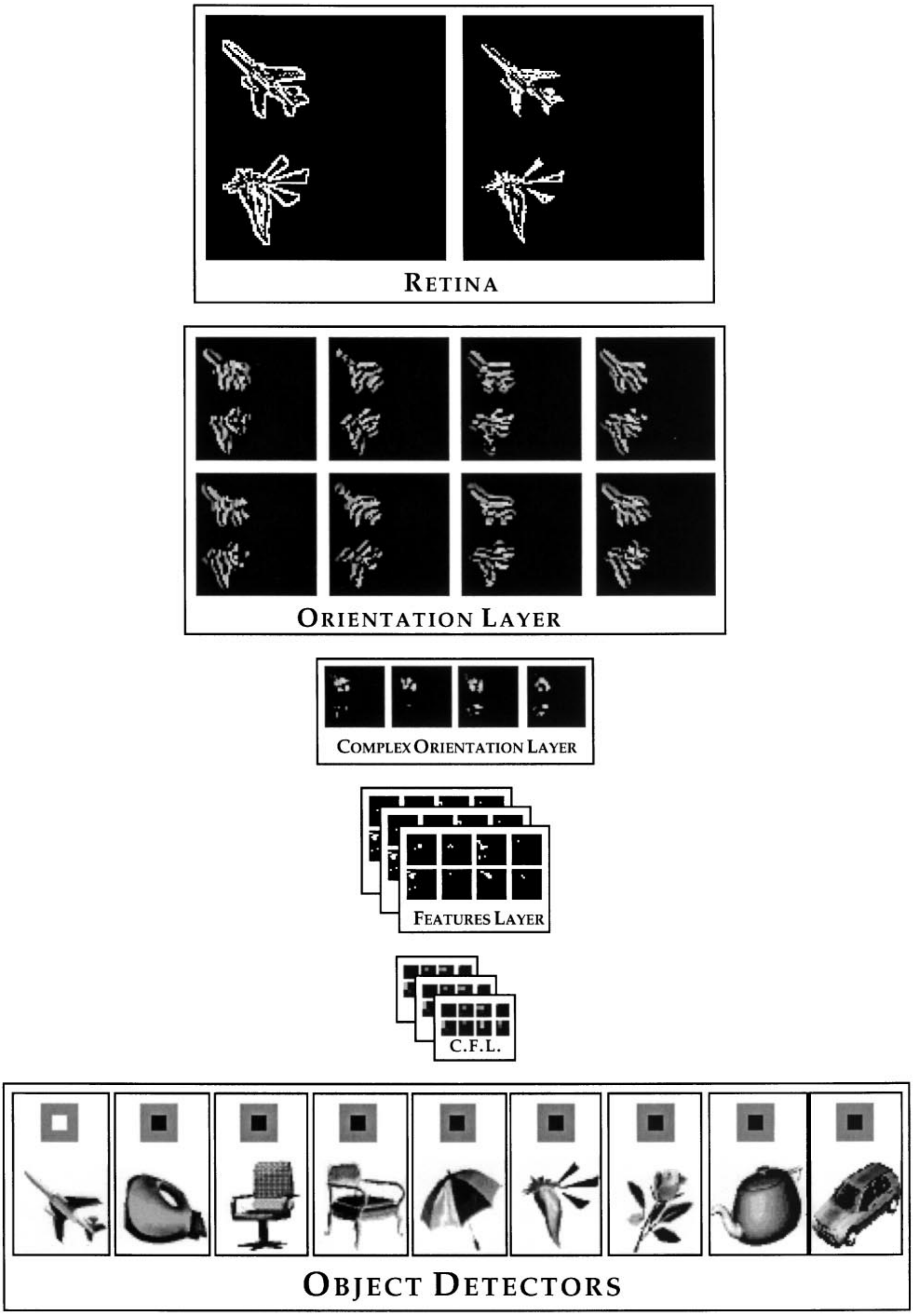


Fig. 3. The 2nd network after propagation of the same input image as in Fig. 2, with attention drawn to the upper-left part of the visual field. Here the attended object is correctly detected.

of the "relevant" neurons would be decreased, giving a temporal precedence to the "relevant information", could account for both computational efficiency and limited resources.

### 3.2. Limited resources model, with attention

In the second model, we kept the preceding model's organization, but we introduced an attentional mechanism, involving a threshold decrease for neurons whose receptive fields fell within a particular region of the visual field. An example of the propagation of an image through that network is shown in Fig. 3.

The computation time for that model was still under 1 s, but the level of performance significantly improved. All possible pairs of objects at all possible locations (different for the two objects) were tested with attention "drawn" (i.e. thresholds decreased) to a region containing one of the 2 targets. In 97% of the images, the network detected one of the targets, which was the attended target in 96% of the images. In contrast, a wrong object was selected for only 2% of the images.

These results seem to indicate that our model of attention constitutes an efficient way to overcome the problems arising with the resource limitations of biological visual systems.

## 4. Conclusion

An important feature of our results is that they can only be exhibited in a network of asynchronously spiking neurons. Lowering the thresholds for a given location in a classic artificial neural network, say a perceptron (with thresholded neurons), would be of no advantage. Neurons at this location would simply reach threshold when receiving a lower weighted sum (i.e. a less specific input). Hence they would be less selective, and performance would decrease. At the same time there would be no processing speed-up, because in such a network neurons need to compute the weighted sum of *all their inputs* at each time step before outputting their response.

A further point that distinguishes our model from most of the existing ones is that it is not only relevant to spatial attention, but can also explain other forms of attention, like feature-selective attention: attending selectively to a particular stimulus feature, such as its shape, orientation, or color, can be viewed as a global lowering of the thresholds of neurons encoding that particular feature, irrespective of their spatial location.

From a more biological point of view, the precise mechanism by which some cells thresholds could be selectively lowered remains unclear. It could for example rely on a localized neuromodulators release, that would affect the membrane properties of the neurons at that location. This is clearly not the only possibility, and we wish to leave that question open for further investigation.

As yet there is no direct physiological evidence for a selective lowering of thresholds for neurons with receptive fields in attended parts of the visual field. Nevertheless, it seems clear that giving a temporal precedence to the information in some retinotopic location is a good and simple way to explain spatial attention. Whether it involves a localized lowering of threshold, or rather some sort of preactivation is still an open question that merits direct neurophysiological investigation.

## References

[1] A. Delorme, R. VanRullen, J. Gautrais, S.J. Thorpe, SpikeNET: a simulator for modelling large networks of integrate and fire neurons. Neurocomputing, submitted.

[2] M.C. Mozer, M. Sitton, Computational modeling of spatial attention, in: H. Pashler (Ed.), Attention, 1998, pp. 341–393.

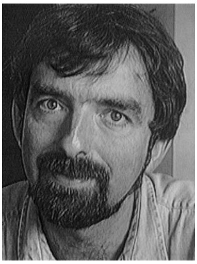
[3] S.J. Thorpe, J. Gautrais, Rapid visual processing using spike asynchrony, in: M.C. Mozer, M.I. Jordan, T. Pesche (Eds.), Neural Information Processing Systems, MIT Press, Cambridge, 1997, pp. 901–907.

[4] S.J. Thorpe, J. Gautrais, Rank order coding: a new coding scheme for rapid processing in neural networks, in: J. Bower (Ed.), Computational Neuroscience: Trends in Research, Plenum Press, New York, 1998, pp. 113–118.

[5] A. Treisman, The binding problem, Current Opinion in Neurobiol. 6 (1996) 171–179.

[6] R. VanRullen, J. Gautrais, A. Delorme, S.J. Thorpe, Face detection using one spike per neurone, Biosystems, 1998, in press.

**Rufin VanRullen** is a Ph.D. student in Cognitive Neuroscience at the Centre de Recherche Cerveau et Cognition in Toulouse, France. His background is in Mathematics and Computer Science. He is currently working on modeling the processes occurring in the primate visual system, e.g. object and face recognition or visual attention. One goal of this work is to explain the astonishing speed of processing in real visual systems when compared to artificial ones. Therefore, his interest has moved towards networks of asynchronously spiking neurons.



**Simon Thorpe** (D.Phil) is a Research Director working for the CNRS at the Centre de Recherche Cerveau and Cognition in Toulouse. He studied Psychology and Physiology at Oxford before obtaining his doctorate with Prof. Edmund Rolls in 1981. He joined Michel Imbert's group in Paris in 1982 and moved to Toulouse in 1993. He has used a range of techniques including single unit recording in awake monkeys, as well as ERP and fMRI studies in humans to study the brain mechanisms underlying visual processing.