INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

# Feature Binding through Synchronized Neuronal Oscillations: A Preliminary Study

Ruggero Milanese

TR-94-044

August 1994

**Abstract**

In this report we analyze the feature binding problem, a combinatorial complexity problem that affects connectionist networks using multiple topographic representations of an image. Inspired from some evidence about the human visual system, we suggest that a solution to this problem may derive by the combined use of attention mechanisms and by exploiting the temporal synchrony of neuronal firing. To this end, a new framework is proposed in terms of a neuronal model, and of a computational architecture capable of producing synchronized firing in distributed assemblies of neurons. This synchronized behavior only affects neurons selected by the network to represent objects of interest. The architecture is structured into a set of feature, conspicuity, and saliency maps, whose neurons are connected in a feedback loop. A number of mechanisms are proposed in order to implement each of these stages, including strategies for reinforcing the synchronous firing of the selected neurons.

# 1 The binding problem

In most applications of neural networks to vision and image processing, such as object classification and texture analysis, a number of representations are constructed from the input image, through convolution with a filter bank. These representations, often called *feature maps*, usually represent the hidden layer of a structured network. If the shape of the target objects is known, a matched filter approach can be used to define the receptive field of feature map neurons, otherwise, banks of oriented spatio/temporal filters can be employed. Figure 1 (left) shows this typical situation, in which the input image is decomposed into three feature maps, representing information about local oriented discontinuities, color and motion.

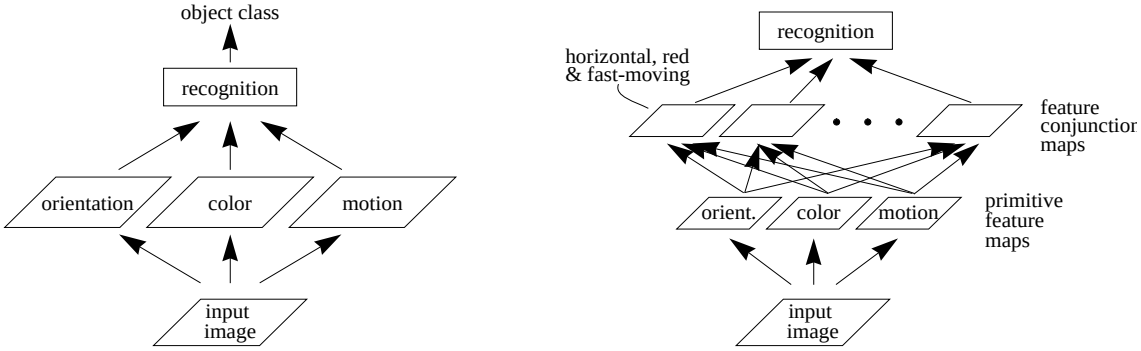


Figure 1: Examples of distributed visual representations. (Left) the recognition layer receives direct input from the feature maps; (right) intermediate conjunction maps represent all possible combinations of values from all feature maps.

This organization is similar to the one found in the primate visual cortex, in which a number of areas have been found, whose neurons have different functional properties. In a similar way to artificial neural networks, most cortical areas appear to use retinotopic coding, i.e. a topographic representation that respects (up to some possible geometric distortions) the topology of the retina.

The role of the third, output layer in artificial neural networks is to recognize objects using the features extracted at the lower level. In this layer, spatiotopic representations are usually abandoned, in favor of more "global" representations that allow for the recognition of multiple objects (in whatever position in the image). This is somewhat similar to the properties found in the *inferotemporal* cortex (IT), which is the most likely biological counterpart of the recognition layer. In this area, receptive field sizes cover almost the whole field of view, and seem to lack any retinotopic coding.

The spatial coding used in these multiple feature maps poses a serious computational problem for the recognition stage. This is known as the *binding problem*, and it consists of the difficulty to represent objects defined through conjunctions of features. Conjunctive coding is essential for representing objects of any practical interests, unless complex, special–purpose feature maps are defined, one for each target object. If general–purpose feature maps are available, as is usually the rule, it is thus necessary to represent particular combinations of them that may be useful to characterize objects. Consider, for instance,

the example of figure 1, and suppose that the object to be recognized is a *car*, defined as something that lies horizontally, is red-colored, and moves fast. Each of these characteristics can be represented by a particular value in the orientation, color, and motion feature maps. Suppose that each of these features is quantized into $D$ discrete values. It is clear that, if there are $L$ feature maps, there are $D^L$ possible combinations of feature values. If this information is to be represented in a retinotopic way, then $D^L$ different conjunction feature maps are required (cf. figure 1, right).

This exponential space complexity poses a serious limitation in the number of feature maps that can be built in practical implementations. Also, current estimates of the number of feature maps present in the visual cortex (at least a dozen) make it unlikely that there are neurons enough to encode each combination of feature. On the other hand, what happens if conjunction feature maps are not used? In this case, the presence of a particular combination of features can still be detected through neurons in the recognition layer. For instance, this could be done using sigma-pi units integrating information over the whole image ($\Sigma$), over a specific combination of feature maps ($\Pi$). However, all spatial information is lost at this stage, and therefore it is not possible to assume any particular spatial layout of this information. This can give rise to many problems, an example of which is the phenomenon of illusory conjunctions. Figure 2 illustrates this problem when the input image contains two objects (a pyramid and a sphere), and two feature maps are extracted, i.e. line segments and circular arcs.

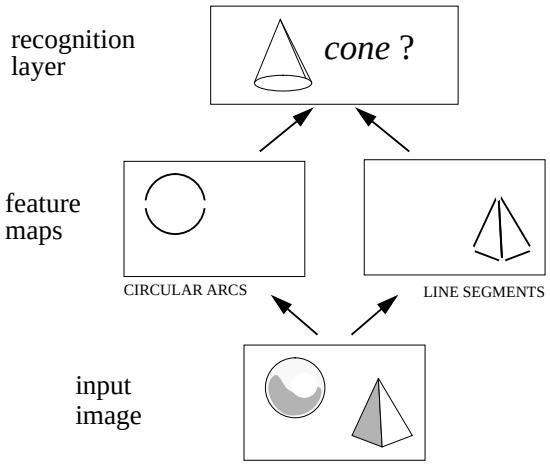


Figure 2: Example of visual illusions generated by the lack of feature binding.

In order to detect the sphere, neurons in the recognition layer should check for the presence of circular arcs, whereas line segments should suffice for the pyramid. However, if the network is to recognize also a third object, a cone, this has to be done by verifying the presence of both circular arcs and line segments. If no binding is performed, arcs and segments would be erroneously combined, suggesting the presence of a cone, even though those features are spatially distinct, and belong to different objects.

It is interesting to note that, in the visual cortex, some conjunctive properties are actually explicitely encoded. In area V4, in particular, neurons appear to be selective to specific combinations of features extracted by earlier stages. Most likely, these combination

are those that define objects of relevance in the subject's everyday life. This allows to detect them very rapidly and with virtually no effort. To some extent, explicit representations of new feature conjunctions appear also to be learnable after some long-term training. However, even with no training, novel objects defined by unprecedented combinations of features are still clearly detectable. This means that some binding across separate feature maps must still be performed, at least to allow the system to detect complex combinations that are not encountered in a systematic way.

So, how can this problem be solved (in both biological and artificial networks) without facing the combinatorial complexity of the conjunctive feature maps? Evidence about the primate visual system suggests that two mechanisms may be useful to this end. The first one is visual attention, i.e. the capability to select a particular portion of the input image, where more processing resources can be deployed in order to solve difficult problems, such as recognition and manipulation. The other mechanism is the dynamic temporal aspect of neuronal activity. The information about which neurons from different maps should be combined to form actual conjunctions may be carried by their temporal pattern of activation. This hypothesis is consistent with the discovery of oscillatory dynamics in the activity of some neurons in the visual cortex. These mechanisms are introduced in the next two sections. In the rest of this paper, a framework is proposed for their integration, together with some implementation details.

## 2 The role of visual attention

Visual attention is a fast, covert mechanism that appears to single out a number of regions in an image, each of which can possibly become the target of an overt, slower saccade. In each region selected by covert attention, visual processing is affected in a number of ways. From a neurophysiological viewpoint, attended neurons provide stronger responses or, if the relevant stimulus is not within their receptive field, their responses are strongly attenuated [24]. On the psychophysical side, a number of studies have shown higher performances (for instance, shorter reaction times) if attention is allowed to move (or is artificially drawn) to a relevant region of the visual field.

Several metaphores have been used to describe these facilitatory effects on selected parts of the input, in terms of a variable zoom–lens [7], or of a spotlight mechanism [14] [32] [36] that can be moved about the retinal image to highlight some parts of it (see [22] for a more extensive survey). In any case, attention appears to be a necessary mechanism for the brain in order to perform some basic functions. A computational analysis of the complexity of the recognition task, for instance, reveals the need for a sequential selection process, since there are not enough neurons to perform it in parallel [22] [33]. Besides the limitation of processing resources, visual attention is also required by some direct links that exist between the visual and motor cortices. For actions requiring visuo-motor coordination, there is the need to continually monitor the relevant part of the input (for instance, an object to be grasped), in order to observe possible changes, and to update movements accordingly. This must be done while the rest of the system keeps processing the whole visual field.

What type of information is used to select the regions of attention? There is evidence that this information ranges from very low–level representations (such as brightness, color, etc.) to high–level ones, related to more cognitive functions. These multiple attention cues

appear to converge in a single center, the superior colliculus (SC), where topographic maps have been found, that receive projections from almost all areas of the visual cortex, as well as from the motor cortex. An important characteristic of these topographic maps is that they are in spatial register, i.e. spatially overlapping, with a limited synaptic distance between neurons of different maps which refer to the same spatial position. These properties make the SC a very probable candidate for the integration of multiple attentional cues. Regions that are salient according to features computed in different areas converge in SC, where topographic lateral connections between different maps could easily allow the selection of a single attention region, perhaps the one receiving the most support across different maps.

The superior colliculus is directly connected to the pulvinar nucleus of the thalamus, which in turn projects back to the whole visual cortex. This feedback is likely to mediate the facilitatory effect observed on the selected region of attention. Several mechanisms have been suggested for the feedback, such as excitatory connections to neurons of the selected area, inhibition of the non–selected neurons, routing of the selected region to the higher cortical areas [1], and generation of highly–correlated spike trains [4].

It should be clear that this attentional mechanism could provide a direct solution to the binding problem. By selecting a small region of the visual field, which likely contains a single object (or even a part of it), higher recognition centers such as IT are relieved from the problem of finding appropriate spatial conjunctions among features. Figure 3 shows in a simplified way how the problem described in the previous section can be solved whenever the attention region containing a single object is used as a gating signal.

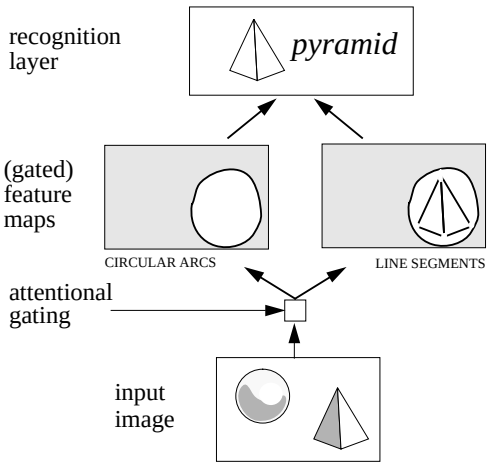


Figure 3: Solution to the binding problem by attentional gating of feature maps.

## 2.1 A computational model of visual attention

In this section, a computational model of visual attention that has previously been proposed is briefly described (see [22] [23] for more details). This model combines bottom–up and top–down cues into a single representation called the saliency map. The integration of these cues is performed through a relaxation process that iteratively modifies multiple measures of interest obtained for each location according to different criteria. Most related to the

scope of this article is the bottom–up component of the system (cf. figure 4), structured on three main levels: the feature maps, the conspicuity maps (CMaps) and the saliency map.

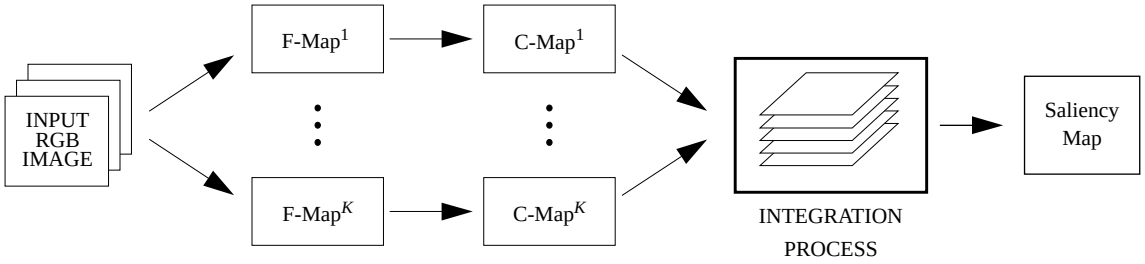


Figure 4: Structure of the bottom–up component of the attention system.

The role of the feature maps (examples are the orientation, edge magnitude, color contrast, and curvature maps) is to represent the image under a number of dimensions, all of which have been shown to be critical to the selection of attention regions. The conspicuity maps are obtained through center–surround operators which extract a measure of interest for each location, based on how different each feature is with respect to the rest of the image. These multiple measures of interest are finally integrated into a single saliency map through a relaxation process which modifies the values of the CMaps according to local inter–map and intra–map connections, until a small number of highly salient regions are extracted.

## 2.2 Open problems

The model described above has been successfully applied to a variety of images, including synthetic images used in typical psychophysical tests, and real images containing multiple complex objects on textured backgrounds. However, some aspects of the system are not fully satisfactory as a solution to the binding problem, and several improvements are required.

The first issue concerns time: for visual search tasks, it has been shown that the amount of time required for the detection of a target may be as low as 40-50 ms.[1] Given that the time scale for neuronal firing is around 10 ms, this means that a very few consecutive computational steps must be sufficient for this process. The model introduced above requires 3-4 steps for the computation of the feature and conspicuity maps, plus a number of iterations for the relaxation process, usually within a dozen. Although it is relatively fast compared to previous approaches (for instance [25]), its time scales are not yet fully plausible.

Another limitation of the system is due to the properties of the relaxation process. As a general rule, the relaxation process modifies the values of the CMaps so as to minimize, among others, an energy measure proportional to the total surface of the selected regions. This is to avoid that these regions grow until they cover the whole image. If the target object is clearly segregated from a regularly textured background, this rule provides a useful constraint, and the results are generally satisfactory. However, if the object is complex and structured, or if there is no clear figure–ground segregation, the contours of the attention

[1] This time may eventually be increased by the presence of distractors, if the target is defined by a unique conjunction of features. Neurons in the IT cortex may even respond to their relevant stimulus in shorter time.

region are often incorrect, comprising only a part of the object. Other problems may appear when the image contains multiple objects that stand out of the background. In these cases, the surfaces of all objects sum up, reducing even further the area of each region. There is thus a limit in the number of objects that can be detected, which usually cannot exceed 2 or 3. Figure 5 contains some examples that illustrate these limitations.

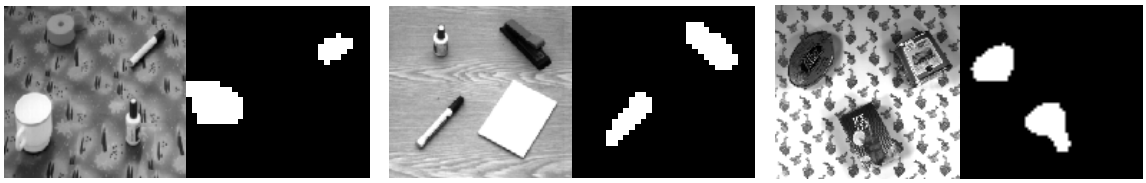

Figure 5: Results obtained by the bottom–up attention system on some difficult images.

Not much can be gained on these issues by simply changing parameters or by slightly modifying the relaxation process. A more powerful way to overcome these limitations could be the introduction of an additional dimension to the spatial representations introduced above, and precisely *time*. The idea that activity in the CMaps and in the saliency map could be oscillatory would add considerable representation power to the system. In particular, activity may continuously evolve through a number of phases, each of which could be assigned to a different target object. At each time instant, only one object would thus have to be selected, so that the relaxation process would only have to isolate it from the rest of the image, avoiding cross–talk from other objects. An advantage of this framework is also that, within each phase, relaxation could be faster, and only one or two iterations may be sufficient to extract the object's shape. Figure 6 shows a sequence of saliency maps that would ideally be obtained on the image shown in the center of figure 5. All four target objects may already be roughly detected during the first period, although at different phases. The following oscillation periods may allow for some further refinements, i.e. for further suppression of noise and cross–talk.

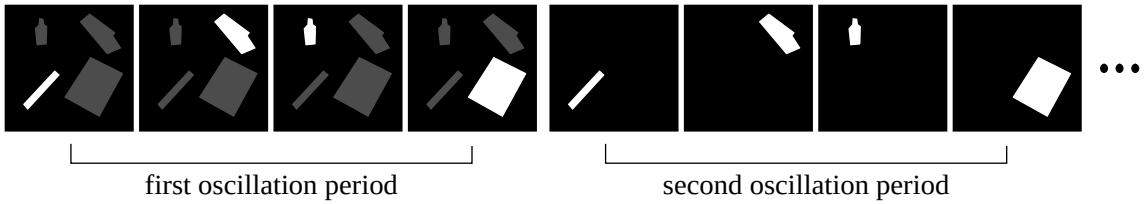


Figure 6: Schematic representation of the temporal evolution of an oscillating saliency map.

## 3 Neuronal oscillations

Oscillatory dynamics have been suggested in section 1 as a possible mechanism for solving the binding problem, by separating in time the firing of cells representing different objects. This mechanism may also be used to improve the performance of the attention system, in particular the integration of multiple conspicuity maps. Below, we briefly review some

fundamental literature in both the biological and engineering fields related to neuronal oscillations (for more extensive reviews, see [30] [16] [28]).

### 3.1 The biology of neuronal oscillations

Oscillatory dynamics in the brain have long been observed and analyzed, for instance through the electroencephalogram. Oscillations appear to take place around a number of central frequencies, from a few Hz (so called $\delta$ range), up to 40 Hz (also called $\gamma$ range). Most interesting are the latter, high–frequency oscillations, since they have been recorded in relation to sensory stimuli (visual, acoustic and olphactory) and to attention.

The first most intriguing results have been recorded in a series of experiments on the visual cortex of the anesthetized cat [5] [9] [10], and later extended to the awake cat and to primates. In these experiments, not only have oscillations in the $\gamma$ range been reported, but their particular phases have been shown to be meaningful. Multiple visual stimuli (for instance, bars) appear to induce synchronized oscillations in the corresponding neurons, when the stimuli, according to common Gestalt laws, appear to constitute a single object. On the other hand, neurons would fail to be synchronized if they respond to perceptually–different objects. A typical example is the "common fate" rule for two moving bars: if two spatially separated, similarly–oriented bars move in the same direction, they would cause synchronous oscillations in the corresponding neurons; whereas if the bars move to different directions, their neurons would not be in phase. In addition, same–phase oscillations have also been recorded between groups of neurons located in different cortical areas, and even in different hemispheres.

These findings suggest that the phase of oscillations may be the "temporal code" that identifies an object as a whole, and allows to keep its featural information separate from that of other objects. This is precisely what would be required to solve the binding problem (cf. section 1). However, there are a number of open issues. One of them is that, within 40 Hz oscillations, it is difficult to keep a large number of separate phases. This poses a limit to the number of objects whose representation can be kept active at the same time. It is thus possible that these synchronized oscillations be linked to the visual attention system, in that they would only affect stimuli that have been attentively selected.

The links between attention and neuronal oscillations may also be useful to explain why stimulus–related oscillations would sometimes fail to be recorded. First, synchronized oscillations would only concern neurons that respond to a very limited part of the visual field, thereby producing weak signals for recording. Second, these oscillations would be induced only when the task requires strong attentive efforts, such as the analysis of a complex image containing multiple objects, or a task involving visuomotor coordination. If the task is straightforward, such as the in the analysis of a well–known, simple object on a uniform background, then it can probably be carried out without the need for oscillations over a large number of periods. It is interesting to note that oscillations can disappear even for visual stimuli that would normally have produced them, if the observer has been subject to intensive training. This suggests that long–term synchronized firings may mediate stimulus–dependent modifications of connections in the cortex (synaptic plasticity), by building a compact "conjunctive representation" of the stimulus, probably at the level of area V4. Once this new representation of the stimulus is built, then it can be accessed directly from

the feature maps, thereby reducing the need for sustained oscillations.

It should be noted that, for the mechanisms described above, there is no need for precise frequency–locked oscillatory dynamics. In fact, what is crucial is the fact that neurons belonging to the same stimulus fire synchronously, whereas neurons belonging to different objects are asynchronous. One one hand, as suggested in [30], oscillations may just occur as a consequence of synchrony. On the other hand, it may simply be easier to produce synchronization from regular periodic patterns rather than from aperiodic ones. In any case, it remains to establish what mechanisms allow the generation of synchronization and/or oscillations. Some hypothesis, involving cortico–cortico connections, have recently been suggested and simulated with artificial neural networks (cf. next section).

### 3.2 Oscillations in neural networks

Independently from the discovery of visually–evoked 40 Hz oscillations in the brain, oscillatory dynamics had already been introduced in artificial neural networks to solve difficult tasks, such as the problem of segregating auditory patterns originated from multiple speakers [20] and the possibility to perform reflexive reasoning using connectionist networks [29]. Interest in these topics has recently increased considerably, in particular with the purpose of developing computational models able to reproduce the neurobiological results.

Several models of neurons have been proposed, that can produce oscillatory behaviors. The simplest ones are mathematical abstractions that consider each neuron as an *oscillator*, which can be uniquely described at each time by its amplitude, phase, and frequency parameters (see for instance [15] [2]). The advantage of this representation is that it makes it easier to mathematically analyze the behaviors of multiple coupled neurons. Another class of models that allow the generation of oscillations is based on assemblies of two coupled neurons forming *dipoles* (cf. figure 7, left). The excitatory neuron $u$ receives input from the sensor, and has excitatory links to itself as well as to the inhibitory interneuron $u$ (excitatory links are represented by the "$\rightarrow$" arrow, whereas inhibitory ones are represented by "$-\circ$").

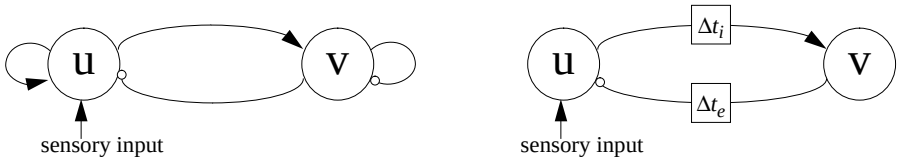


Figure 7: Different schemes of dipole connections able to generate oscillations.

When the dipole receives a steady input of sufficient amplitude, the activity of neuron $u$ increases, also thanks to the recurrent excitation. However, the activity of the $v$ neuron also increases, until its value is high enough to inhibit $v$ and to reset its potential. The activity of the $u$ unit will then again increase, thanks to the sustained input. The differential equations defining the dynamics of the dipole require the specification of several parameters, which must be carefully chosen in order to obtain the desired behavior (cf. [21] for more details).

Another way to generate oscillatory dynamics is by introducing temporal delays in the connections between the excitatory and inhibitory neurons (cf. figure 7, right). It is clear that, while the activity of the $u$ unit increases thanks to the input, the same increase will be

propagated to the unit $v$ after a time delay $\Delta t$, thereby resetting the activity of the $u$ unit to zero. Delays in synaptic transmission are actually quite plausible in biological hardware, but, again, care must be taken in the choice of the parameters, especially in the amount of delay.

Apart from the capability to generate oscillations, it is important that a neuronal model can be used to synchronize the activity of multiple units. This must be obtained through appropriate couplings between units. It is easier to describe some of these couplings for the case of formal, abstract neuron models. For the sake of simplicity, the frequency of oscillators will be considered constant, whereas the amplitude will generally be dependent on a function of the input. Let thus $\phi_i(t)$ be the istantaneous phase of unit $n_i$. A simple connection scheme is the *nearest neighbor* coupling, defined in 1-D as follows [15] [11]:

$$\frac{\partial \phi_i(t)}{\partial t} = f(\phi_{i-1}(t)) + f(\phi_{i+1}(t)), \tag{1}$$

where $f$ is a generic odd function. Even with a large number of iterations, this model has difficulties in synchronizing units, especially if there are gradients in the input signal. A better coupling model has been shown to be the *comparator*, in which all units are connected to a single unit, which in turn provides feedback to each of them through their average value:

$$\frac{\partial \phi_i(t)}{\partial t} = f(\phi_i(t) - \frac{1}{N}\sum_{j=0}^{N-1}\phi_j(t)). \tag{2}$$

However, even in this case, if units receive strongly different inputs, they may lose coherence and decouple from each other [15]. Other interesting coupling schemes have been suggested in [11]. One of them is a refinement of the previous one, in which the units feed to multiple comparator units $Z_i$, which in turn provide feedback for the updating of the $\phi_i$ values. This model, called *adaptive filtering* coupling, is defined as follows:

$$\begin{aligned} \frac{\partial Z_i(t)}{\partial t} &= f(\frac{1}{N_1}\sum_{n=-N_1/2}^{N_1/2}\phi_{i+n}(t)) \\ \frac{\partial \phi_i(t)}{\partial t} &= f(\phi_i(t) - \frac{1}{N_2}\sum_{n=-N_2/2}^{N_2/2}Z_{i+n}(t)), \end{aligned} \tag{3}$$

where $N_1, N_2$ define the number of input, respectively for the $Z_i$ and $\phi_i$ units. Finally, a model called *cooperative bipole* coupling is defined through a convolution of the units activities with a weight function $w(n)$:

$$\frac{\partial \phi_i(t)}{\partial t} = f(\sum_n \phi_n(t)\cdot w(n-i)). \tag{4}$$

Its name is due to the characteristic"$\infty$" shape defining the support of the function $w(n)$ (where the index $n$ characterizes a 2-D position).

All previous coupling schemes can readily be implemented using both the dipole and the delayed–connection models. The nearest neighbor coupling, for instance, can be obtained by

simply establishing associative, excitatory links between the excitatory unit $u_i$ and the units $u_{i-1}, u_{i+1}$ of neighboring dipoles (cf. figure 8, left). An empirical demostration of phase locking properties of this scheme is given in [34]. More global connection schemes can be obtained by simply increasing the number of these associative connections for each dipole. Another extension of this connection scheme allows the possibility to either synchronize or desynchronize dipoles. This is done by adding excitatory connections to a common inhibitor unit $v$ [21] (cf. figure 8, right). It has been shown that, if the strength of the connections to the $v$ unit is high enough, the two dipoles become desynchronized, i.e. their coefficient of correlation is very low. If instead these connections are weak, then the network behaves as in the previous case, i.e. the dipoles synchronize their activity.

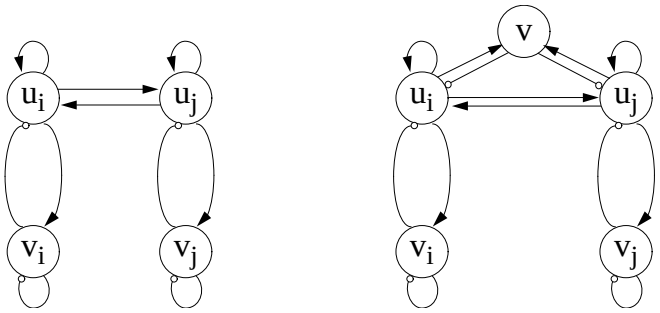


Figure 8: Connection schemes allowing synchronization of oscillating dipoles.

In the case of two dipoles $i, j$ with delay—induced oscillations, synchronization can be obtained through excitatory links of the type $u_i \rightarrow v_j$ and $u_j \rightarrow v_i$, each of which be affected by some time delay $\Delta t_{uv}$.

In addition to abstract oscillators and oscillating dipoles, other neuronal models exist which can produce synchronization. One of them is a refined version of the sigma–pi unit, which converts the input signal into spike trains [6]. This unit has two types of inputs, provided by *feeding* and *linking* synapses. The first ones represent the main, sensory–driven signal path, whereas the second ones act as amplifier, causing a shift in the phase of the output pulses. A similar solution has been proposed, although in less detailed way, in [13]. Another model that retains the capability to synchronize the firing of units without requiring oscillations has been proposed in terms of simple McCulloch–Pitts neurons [17]. If these neurons are connected in an excitatory all–to–all fashion with a global inhibition, they synchronize their bursts whenever a sufficient number of input signals coincide. It is interesting that, as a consequence of this synchronization, oscillations in the network activity tend to emerge spontaneously.

The models described above focus on the possibility to obtain synchronization or desynchronization as a consequence of the network topology and of the input signal. Some work has also been done recently to show how these mechanisms can be used to solve practical problems. One of them is the problem of texture segregation, for which an architecture has been proposed, in which the input image is first convolved with a bank of Gabor filters [2] (in a similar way to [19]). The responses of the convolution with these filters can be used as the amplitude values for a corresponding bank of oscillators. The phases of the oscillators interact by means of connections between each unit and some neighboring units within the

same map. This is thus similar to the adaptive filtering coupling scheme described by equation 3. However, the strength of the connections are proportional to the distance between the cells, as well as to their amplitude. Through this coupling rule, the arrays of oscillators can be able to separate two regions of different textures. However, there is no inter–map coupling between the oscillators. For this reason, only some of the oscillator arrays can effectively discriminate the two regions, and no feature binding – as defined in section 1 – is actually performed. Examples of networks which do contain this kind of lateral inter–map connections can be found in [12] [8].

Another interesting applications of coupled oscillators is in figure–ground segregation. The architecture proposed in [21] consists of multiple 2-D arrays $F^k_{x,y}$ of oscillators, each value of $k$ representing a different feature. Within each map $k$, all oscillators are linked with each other through excitatory connections. This also holds for all units at the same location $x, y$ in different maps. These excitatory connections are balanced by inhibitory ones, similarly to the comparator model described by eq. 2. To this end, three types of "global" inhibitory units are used, with connections from/to: all units within each plane, all units at the same position, and all units in the whole network. This network was tested on an image containing two squares on a uniform background. After a transient period, all oscillators belonging to the same object appear to synchronize with each other, and to keep a constant phase lag with respect to those belonging to the other object.

An array of oscillators has also been applied to the recognition and completion of patterns. It is well known that, by using a Hebbian rule to define connection strengths, a network can be used to store and recognize patterns as an associative memory. In addition, if the units are oscillators, it has been shown that it is possible to recognize multiple patterns, even if they are presented simultaneously [34]. In this case, the assemblies of neurons representing a single pattern oscillate in a phase–locked way, and those representing different patterns have different phases, allowing the network to separate these patterns in time.

Finally, oscillatory neuronal models have also been studied in conjunction with attention mechanisms [26] [27]. In the first of these studies, the firing rate of V1 neurons is modeled with an inhomogeneous Poisson process, with a sinusoidal mean firing rate. The amplitude of the sinusoidal function is assumed to be determined by a measure of interest provided by some external saliency map. In the second of these studies the saliency map is not used as a direct sinusoidal modulation, in order not to affect the mean firing rate of the neurons, which should rather be determined only by the sensory input. Instead, the saliency map provides a signal that forces synchronization in the selected area. This is done by adding a term to the mean firing rate of attended units which instantaneously increases to a high value, and then drops for a longer period of time. Since the integral of this modulatory function over a whole period is 1, the mean firing rate is not affected. In this way, all attended units fire synchronously, and then enter a refractory period which precedes a new series of bursts.

# 4 Outline of the proposed framework

In the previous sections, visual attention and neuronal synchronism have been introduced as possible mechanisms for the solution of several problems in vision, including feature binding,

feature–ground and texture segregation. In this section we briefly summarize some ideas on how these two mechanisms could be combined into a complete vision system, whose structure loosely resembles the one of the human visual system. Figure 9 describes the overall architecture.

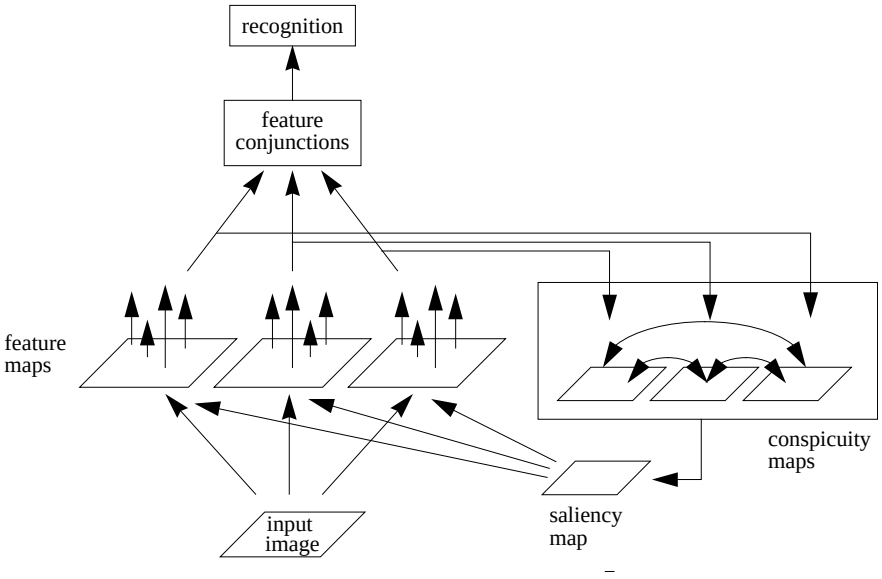


Figure 9: Overview of the proposed framework.

The input image is assumed to be simply a static storage, providing the sensory input to the stage represented by the *feature maps*. These are basically the same maps introduced in the previous architecture (cf. figure 4). However, the transfer of information from the feature maps to the other stages of the system is not done in an instantaneous, synchronous way. Instead, the release of information towards the conspicuity maps and the feature conjunction module is done continuously, each neuron firing with a latency characterized by its own local feature intensity.

The pathway reaching the *conspicuity maps* is part of a feedback loop that has the goal of modifying the temporal properties of the neurons in the feature maps. Neurons of the conspicuity maps compute a measure of interest through center–surround receptive fields, which enhance regions containing different feature values from the rest of the image (cf. section 2.1). Since different conspicuity maps represent different interest measures, an integration process is required, in order to obtain the *saliency map*, i.e. a single binary map identifying the most interesting region of the image. The most likely biological candidates for these stages are the superior colliculus and the pulvinar.

The saliency map, representing the result of the attention system, is used to close the feedback loop with the feature maps. This is done through topographic connections between each neuron of the saliency map and all neurons of the feature maps at the corresponding locations. The role of these connections is to modify the temporal firing pattern of the feature maps. One one hand, the neurons belonging to the selected area will be forced to synchronize throughout all feature maps. On the other hand, the firing pattern of the remaining neurons may be forced to be even more decorrelated. The temporal coherence

of the assemblies characterizing the attended objects will thus continuously increase while the feedback loop is iterated. It should be noted that, while the saliency map is being computed for the selection of one object, the feature maps keep generating spikes. This new information may thus be useful to select a new, different object. So, the dynamics of the whole network evolves according to two main time scales. One is the time scale of the feedback loop "feature maps $\rightarrow$ conspicuity maps $\rightarrow$ saliency map $\rightarrow$ feature map $\rightarrow$ ..." which allows to synchronize neurons belonging to an attended object $A$. The other one involves a certain number of consecutive feedback loops, which allow other objects to be selected in a sequential way, before the neurons of object $A$ fire again synchronously. The time scale of the latter process may be thought of as the period of an oscillatory pattern similar to those observed in the biological systems (cf. section 3.1).

The two final processing stages are the *feature conjunction* and the *recognition* layers. These two components are meant to perform similar functions to those of visual areas V4 and IT, respectively. The first one is assumed to be composed of neurons that can dynamically modify synaptic strengths in their input connections from the feature maps, if highly synchronous signals are received for a long time. In this way new conjunctive representations may be learnt, if the same input pattern is repeatedly presented to the system. The final module performs object recognition, by matching this feature information to a number of stored models, and by taking a final classification decision.

The next sections describe in more details the possible implementation of some of these components, in particular the feature and conspicuity maps, the saliency map, and the feedback loop.

# 5 The neuronal model

As suggested in section 3.2, there are many choices in design of a neuronal model allowing synchronization. On one extreme, the choice of abstract oscillators makes their use simpler, on the other hand, more detailed models, such as the Hodgkin–Huxley one, would be more plausible, and may offer richer behaviors (see for instance [18]).

Models such as the oscillating dipole are somewhat in between, combining a certain level of detail with some intuitive understanding of their dynamics. However, the choice of dipoles is not qualitatively different from abstract oscillators, since it is not necessarily more plausible from the biological point of view. Furthermore, it implies that all units in all maps will act as oscillators, with a regular periodic activity. This may not be a necessary condition. In fact, periodic oscillations may be useful to constrain synchronism for the units belonging to the attended areas, but it may instead be appropriate to have aperiodic behaviors in the neurons that belong to the background, since this may help decorrelate their activity.

For these reasons, we propose a neuronal model that is structurally simpler than the dipole, but which may also offer richer behaviors. The model describes a neuron by representing its firing time through a stochastic function. This is somewhat related to Poisson models of neuronal firing rate, except that in each time unit the output of a neuron is binary, i.e. it either emits a spike or it does not. More precisely, for each neuron $n_i$ we define a stochastic function $p_i(t)$ describing the probability to emit a spike at time $t$. This function is defined through a periodic combination of random deviates, which depends on

three main parameters: the frequency $\omega_i$, the phase $\phi_i$, and the "uncertainty" width $\sigma_i$. For the sake of simplicity, we have decided to implement these stochastic components through Normal (Gaussian) deviates:

$$\begin{aligned} p_i(t) &= Prob(\text{neuron } n_i \text{ spikes at time } t) \\ &= \sum_{n \geq 0} \frac{1}{A_i} \cdot \exp - \frac{(t - \frac{n}{\omega_i} - \phi_i)^2}{2{\sigma_i}^2} \cdot \mathcal{W}\left(2\omega_i(t - \frac{n}{\omega_i} - \phi_i)\right), \end{aligned} \tag{5}$$

where $\mathcal{W}(\xi)$ is a window function that is 0 everywhere, except in the interval $[-1, 1]$, in which it is 1. This means that each Gaussian component is bounded to the interval defined by extrema $\frac{n}{\omega_i} + \phi_i \pm \frac{1}{2\omega_i}$, $n = 1, 2, ....$ The normalization coefficient $A_i$ is defined as $\int \exp - \frac{(t - \frac{n}{\omega_i} - \phi_i)^2}{2{\sigma_i}^2} dt$ over any any of these intervals. It should be noted that $p_i(t)$ is a probability distribution in the strict sense only within each of these intervals of length $1/\omega_i$. However, we shall generally refer to it as the "probabilistic" model of a neuron's firing time.

In order to characterize the firing pattern $f_i(t)$ of a neuron $n_i$, describing the times at which $n_i$ fires, the three parameters $\omega_i$, $\phi_i$, $\sigma_i$ must thus be determined. This can be done in a natural way, by considering some known properties of biological neurons. It is known, for instance that the firing rate of a neuron (or, in other words, its frequency) is determined by the similarity between its input and its preferred stimulus, as determined by the profile of its receptive field. An inverse relation holds for the neuron's *latency*, i.e. the time elapsing from the stimulus presentation until the neuron starts discharging. In this case, input stimuli similar to the optimal one will induce a shorter delay. Finally, it seems reasonable to assume that a non–optimal stimulus will also produce a more noisy temporal pattern in the neuron's spikes. This can be modeled in terms of a higher uncertainty in its firing model, by increasing the width of the stochastic deviates. On the basis of the above, we define the values of the parameters $\omega_i$, $\phi_i$, $\sigma_i$ of a neuron $n_i$ to be a function of the similarity between its current input and its optimal stimulus. Since this similarity is generally obtained by convolving the input image with some filters, we define the convolution results as the "intensity" of a neuron, from which these three parameters can readily be derived. Figure 10 shows a qualitative description of the relations between $\omega$, $\phi$, $\sigma$ and the intensity of a feature map neuron.

As compared to the formal oscillators (as well as to dipoles), these relations between the intensity of a neuron $n_i$ and its probabilistic model of firing $p_i(t)$ clearly allows more complex dynamics in the system. Figure 11 describes the temporal evolution of this model for three neurons characterized by three different levels of intensity.

It can be seen that, due to its small $\sigma_i$, a high–intensity neuron will tend to fire in a periodic way, with very little shift from the values of $t = \frac{n}{\omega_i} + \phi_i$, $n = 0, 1, 2, ...$ This means that, if the power spectrum of its firing pattern $f_i(t)$ is considered, this will be characterized by one isolated peak (in addition to its harmonics, and the dc component). In other words, its auto–correlation function will present a high peak at $1/\omega_i$. Furthermore, if $n_i, n_j$ are two neurons characterized by the same parameters, with a low uncertainty value $\sigma$, then they will be highly synchronized, meaning that the cross–correlation function of their firing patterns will present a high peak at 0.[2] On the other hand, if $n_i$ is a low–intensity neuron,

[2] An analogous way to measure their synchronization is through the correlation coefficient

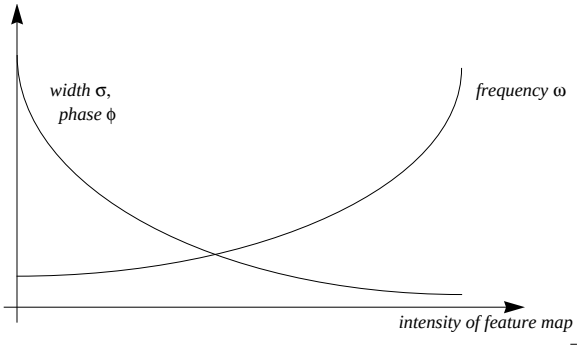


Figure 10: Qualitative description of the relationships between the intensity of a feature map neuron and the frequency, phase, and width parameters of its firing probability $p(t)$.

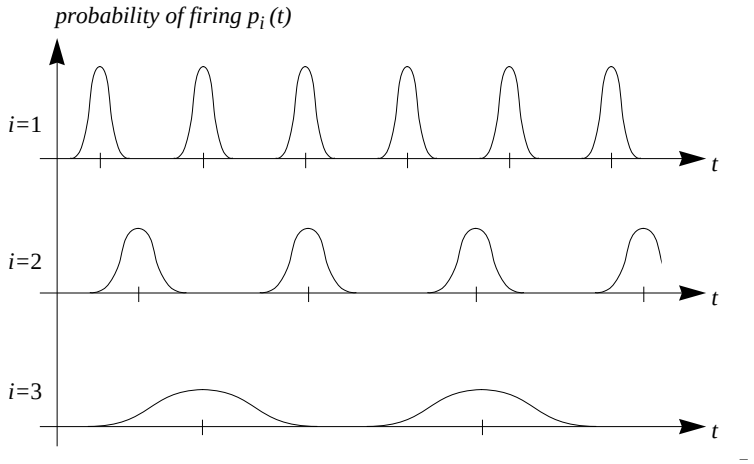


Figure 11: Probabilities of firing for three feature map neurons having different intensities.

its high $\sigma_i$ value will cause a large scatter in its firing time. This can be seen in terms of a flatter power spectrum, as well as a lower value of the cross–correlation function at 0, for any pair of similar neurons.

As described in the previous section, feature map neurons will be affected by the feedback loop from the saliency map. The main goal of this feedback is to increase synchrony between neurons belonging to the attended region, throughout all feature maps. This can be done through topographic connections between saliency map neurons $S_i$ to all feature maps neurons $F_i^k, k = 1, ..., K$. For instance, the scatter of all phases $\{\phi_i^1, ..., \phi_i^K\}$ can be reduced, in a proportional way to the intensity of the $i$-th saliency map neuron. In a similar way, the uncertainty of each neuron's firings over time, i.e. the value of the parameters $\sigma_1^k, ..., \sigma_i^K$ can also be reduced, so that the periodicity of their firing pattern will be increased.

Given that the $\phi$ and $\sigma$ parameters are changed over time through feedback, the frequency of firing (i.e. the value of $\omega$) is the last parameter that can be left to express the original intensity of the neuron, i.e. the similarity of its stimulus to its optimal pattern. If

$\rho_{i,j} = \frac{E[f_1 f_2] - E[f_1]E[f_2]}{\sqrt{E[f_1^2] - E[f_1]^2}\sqrt{E[f_2^2] - E[f_2]^2}}$, where $E[f]$ is the mean value of $f$.

the value of $\omega$ was also to be changed by feedback, then the coding properties of feature map neurons would be completely destroyed. Imagine, for instance, an image from which a green, vertical object has been selected through the saliency map. Suppose, for simplicity, that there are two feature maps, representing vertically–oriented discontinuities, and red–colored pixels. Then, if feedback from the saliency map was allowed to increase the firing frequency of neurons in the selected region, this would have the effect that the object would now be represented as if it was *red*. In order to be able to use the information generated by the feature maps at the higher levels of the system (cf. the conjunction and the recognition layers in figure 9), it is thus important not to completely override their stimulus–related firing frequencies.

Despite the necessity to preserve the coding property of neurons, slight changes in the $\omega$ parameters may be tolerated, and actually be useful. For instance, some diffusion process smoothing the $\omega$ values of the neurons in the selected region may allow for some noise removal and for some filling–in, as found in the visual cortex [10]. Also, some feedback in the frequencies $\omega$ of attended neurons may allow for an increase of their correlation coefficient, by forcing them to approach the closest integer fraction of the highest one. In other words, if the location $i$ belongs to the selected region, and $\omega_i^k \geq \omega_i^h, \forall h \neq k$, then the values of $\omega_i^h$, $h \neq k$ should be forced to approach $n\omega_i^k$, where the integer $n$ is defined by $\text{argmin}_m \mid \omega_i^h - m\omega_i^k \mid$. In this way, neurons of different maps belonging to an attended object will keep firing with different frequencies, all of which are integer multiples of a basic frequency.

The above discussion describes the effect of feedback on the neurons belonging to an attended region, in terms of an increased synchrony in their firing pattern. It is possible to further increase the separation between the behavior of neurons in the selected region and those in the background. This can be done for instance by introducing some noise in the parameters of neurons that have not been selected. An alternative could be to continuously increase at some small rate the values of their $\sigma$ parameter. In this way, the periodic character of their firing times, as well as their correlation coefficient would continuously decrease.

# 6 The feature maps

The neuronal model described in the previous section is the basic unit of the feature maps. In this section we shall first describe how the "intensity" information for feature map neurons is computed, and then explain how this information is turned into initial values of the frequency, phase, and uncertainty information of the $p(t)$ functions describing their firing pattern. In terms of notation, the 1-D index used above to identify neurons will be replaced by an "$x, y$" one, identifying their 2-D position.

## 6.1 Computing feature map intensities

In a previous simulation of the attention system, the following feature maps had been proposed: local orientation, edge magnitude, curvature, and red–green, blue–yellow color opponency maps [22] [23]. In the present context, we propose the use of the same maps, with some modifications in the definition of the orientation one. Rather than merging all

orientation information into a single map, a number of them will be used to indicate the presence of discontinuities for each orientation index. This is done by convolving the input image with a bank of oriented Gaussian 1-st derivative filters:

$$\mathrm{GD}_{x,y}(\vartheta) = A \cdot \frac{x\cos\vartheta + y\sin\vartheta}{\sigma_x} \cdot \exp\frac{-(x\cos\vartheta + y\sin\vartheta)^2}{2\sigma_x^2} \cdot \exp\frac{-(-x\sin\vartheta + y\cos\vartheta)^2}{2\sigma_y^2}, \quad (6)$$

where $A = \frac{-1}{(2\pi\sigma_x\sigma_y)}$, and $\sigma_x, \sigma_y$ define the scale parameters on the two canonical axes. This filter provides a good approximation of the receptive field profile of a large class of V1 cells, and is convolved at $N$ orientations to provide the intensity information for $N$ orientation maps $F_{x,y}^{orient}(\vartheta)$. The orientation values are taken in the interval $[0,\pi]$, i.e. $\vartheta \in \{0, \frac{\pi}{N}, \ldots, \frac{(N-1)\pi}{N}\}$. Directional information is discounted by taking the absolute value of the convolution. A further, edge magnitude feature map is obtained by taking the maximum absolute value of the response throughout all orientations: $F_{x,y}^{magn} = \max_\vartheta \mid I_{x,y} \star \mathrm{GD}_{x,y}(\vartheta) \mid$.

Another differential operator is used to compute the intensity for the curvature feature map: $F_{x,y}^{curv} = \mathrm{div}\left[\frac{\nabla I}{\|\nabla I\|}\right](x,y)$, where $\nabla I$ is the gradient of the input image. Finally, two color–opponency filters with a Gaussian profile were used in [22] to compute the color contrast feature maps. In the present simulations, if no color information is used, these feature maps are replaced by $F_{x,y}^{grey} = I_{x,y}$, which corresponds to the grey–level intensities of the input image.

### 6.2 Computing initial frequency and latency of feature map neurons

The feature map intensities described above must be converted into initial values for the $\omega, \phi, \sigma$ parameters of each neuron's firing model. Only the initial values need to be computed, since their dynamics will be determined by the feedback from the saliency map.

The general rules for the definition of the initial values has been qualitatively described in section 5. Since the intensity values of all feature map $F_{x,y}^k, k = 1, .., K$ are restricted to the interval $[0,1]$, these rules can easily be implemented in the following way:

$$\omega_{x,y}^k(0) = (\omega_{max} - \omega_{min}) \cdot z(F_{x,y}^k) + \omega_{min} \quad (7)$$

$$\phi_{x,y}^k(0) = (\phi_{max} - \phi_{min}) \cdot z(1 - F_{x,y}^k) + \phi_{min} \quad (8)$$

$$\sigma_{x,y}^k(0) = (\sigma_{max} - \sigma_{min}) \cdot z(1 - F_{x,y}^k) + \sigma_{min}, \quad (9)$$

where $\phi_{min}, \phi_{max}, \omega_{min}, \omega_{max}, \sigma_{min}, \sigma_{max}$ define lower and upper bounds for these three parameters. The function $z(\xi) : [0,1] \to [0,1]$ may be simply the identity function, for a uniform quantization of the intensity values. Alternatively, it may be useful to better discriminate between high–intensity values or low–intensity ones, in which cases, the functions $z(\xi) = \xi^2$ and $z(\xi) = \sqrt{\xi}$ may for instance be used.

The choice of the initial values of $\omega, \phi, \sigma$ through the above rules has some nice properties in terms of facilitating the segmentation process and the synchronization of objects of interest, whenever these objects are characterized by some uniformity in at least some of their features. This smoothness constraint is a generally accepted assumption, and is at the

basis of most pattern recognition techniques. Since the initial values of these parameters depend on the feature map intensity values, an object with uniform features will have the same firing model for all its neurons. This means that, if the object is uniform for a specific feature map, its neurons in this feature map will be already synchronized from the very beginning. This is not true, however, within other feature maps, and for assemblies of neurons belonging to different feature maps (this is the reason why feedback from the saliency map is required).

The possibility that neurons of some object will be initially synchronized within a certain feature map may appear to be arbitrary and biologically not plausible. After all, a considerable number of synapses is required for the sensory information collected by the retina to reach the cortical areas that represent these feature maps. However, there is multiple evidence supporting the idea of precise timing in the information that reaches the cortex. This is confirmed by some known visual illusions that can be obtained by artificially changing the luminance of two stimuli.[3] More recently, precise neuronal timings have been shown to be essential in order to explain the capability of the visual cortex to recognize known stimuli in very short times. As suggested in [31], one or two spikes may be sufficient for each neuron in the pathway from the retina to area IT, in order to recognize a relevant pattern. The rest of the time is probably used just to refine processing and to resolve ambiguities, through feedback from the activated IT neurons back to the whole pathway.

# 7 The conspicuity maps

The feature maps described in the previous section contain populations of neurons, each of which generates spikes according to its probability model $p_i(t)$. These spikes represent the input to the neurons in the conspicuity maps, whose goal is to select, from each feature map, the locations containing the most salient information. In [22] [23] a strategy has been proposed for the calculation of these maps, for the case of case of instantaneous, synchronous processing. Briefly, this is done through a bank of multi–scale, multi–orientation filters defined by the difference of oriented Gaussians:

$$\mathrm{DG}_{x,y}(\vartheta,\varsigma) = A_1 \cdot g_{x,y}(\vartheta,\varsigma,r_1\varsigma) - A_2 \cdot g_{x,y}(\vartheta,r_2\varsigma,r_1r_2\varsigma), \tag{10}$$

where $r_1, r_2$ are fixed constants defining scale ratios, and the functions $g_{x,y}(\sigma_x,\sigma_y)$ are 2-D Gaussians, with scale factors $\sigma_x,\sigma_y$ on the two canonical axes. The constants $A_1, A_2$ are defined so that the integral of each component of the DG filter is 1 over the finite support given by the image size. These filters are computed over a number of scales $\varsigma_1,..,\varsigma_M$ and at different orientations, at the same values of $\vartheta$ used in the definition of the GD filters (cf. eq. 6). For each position $x, y$ the conspicuity maps $C^k_{x,y}$ are obtained by taking the maximum response of the convolution of $\mathrm{DG}_{x,y}(\vartheta,\varsigma)$ with feature map $F^k_{x,y}$, over all values of $\vartheta,\varsigma$.

[3] In the Hess effect, for instance, the two stimuli are moving at the same speed, but the difference in luminance causes an increase in latency that makes one stimulus appear as if it were lagging the other [35] [3]. Another illusion, called the Pulfrich effect, is obtained by viewing a swinging pendulum with a light–attenuating filter before only one eye [3]. In this case, the increased latency on the filtered eye causes a change in the perceived depth, which makes the pendulum appear as moving in an elliptical path.

Although this strategy is still applicable to the present context, some changes must be made in the way it is implemented. Due to the continuous generation of spikes by the feature map neurons, the activity of neurons in the conspicuity maps must also be evaluated over time. It is thus necessary to discretize time, using some temporal interval $\tau$. This constant must be small enough to allow for precise evaluation of the neuron's dynamics: if it is too large, then all the computation becomes synchronous and it is not possible to exploit any temporal structure in the neuron's firing. The value of $\tau$ also represents the time interval in which neurons integrate their inputs. So, $\tau$ must be large enough that neurons in the conspicuity maps can receive enough excitation for actually generating spikes.

Given the time constant $\tau$, and the firing probability model of feature map neurons, we must determine the information reaching neurons in the conspicuity maps for each time interval $[n\tau,(n+1)\tau]\, n \geq 0$. In this way, their activity can be determined through the filtering procedure described above. This can be done in several ways. One possibility is to define the variable $\mathcal{F}^k_{x,y}(n\tau)$ describing the amount of spikes generated by neuron at location $x, y$ of feature map $k$ within the $n$-th time interval, and to convolve the maps of $\mathcal{F}^k_{x,y}(n\tau)$ values with the filter banks defined by eq. 10. The $\mathcal{F}^k_{x,y}(n\tau)$ values can be either computed by counting the number of events generated by the stochastic model $p^k_{x,y}(t)$ in the time interval $[n\tau,(n+1)\tau]$, or they can more efficiently be approximated through their average value (cf. figure 12):

$$\mathcal{F}^k_{x,y}(n\tau) = \int_{t=n\tau}^{(n+1)\tau} p^k_{x,y}(t)dt. \tag{11}$$

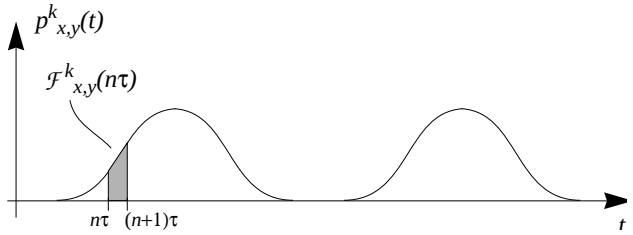


Figure 12: Computation of the input $\mathcal{F}^k_{x,y}(n\tau)$ to neurons in the conspicuity maps through the average value of the firing probability $p^k_{x,y}(t)$ in the interval $[n\tau,(n+1)\tau]$.

An alternative is to consider the model $p^k_{x,y}(t)$ as a *mean firing rate*, and to define $\mathcal{F}^k_{x,y}$ through a stochastic process $P$ which is a function of $p^k_{x,y}(t)$. One example for such a $P$ process is the Poisson distribution (cf. [26] [27]).

At each (discretized) time $n\tau$, we define the "intensity" measure $C^k_{x,y}(n\tau)$ for the neurons of the conspicuity maps through the convolution of the maps $\mathcal{F}^k_{x,y}$ described above with the bank of DG filters:

$$C^k_{x,y}(n\tau) = \max_{i,j}\ \mathcal{F}^k_{x,y}(n\tau) \star \mathrm{DG}(x,y,\vartheta_i,\varsigma_j). \tag{12}$$

# 8 The saliency map

In section 4 the saliency map was introduced as the layer of neurons that receives information from the conspicuity maps in order to build an overall binary mask representing the objects of interest. In [22] [23] a method was proposed for the construction of the saliency map, based on a relaxation process run over the conspicuity maps (cf. section 2.1). In the present context this process, based on an iterative energy–minimization procedure, may not be necessary. The information computed by the feature maps is not transfered to the conspicuity maps in an instantaneous way, but is rather distributed throughout a large temporal interval. For this reason, most of the time the values of $\mathcal{F}^k_{x,y}(n\tau)$ will be very low, causing very little activity in the conspicuity map neurons. However, this activity will rapidly increase to its highest value when $n\tau$ approaches one of the peaks of $p^k_{x,y}(t)$, located at $t = \phi^k_{x,y} + \frac{m}{w^k_{x,y}}, m = 0, 1, 2, ...$ With respect to the synchronous framework, the selection task performed by the conspicuity maps is thus much simplified. For the same reason, the problem of merging them into the saliency map is also easier, since it is very unlikely that at one time step that there will be more than one map capable of selecting conspicuous objects. It may thus be possible to integrate the conspicuity maps by simply defining the intensity of the saliency map neurons through a weighted sum:

$$S_{x,y}(n\tau) = \frac{1}{K} \sum_{k=0}^{K} w_k \cdot C^k_{x,y}(n\tau), \tag{13}$$

where the weights $w_k \in [0, 1]$ may either be constant or, if some a–priori knowledge is available, may favor features that are known to characterize a target object. However, there is still the possibility that the same conspicuity map may detect multiple objects at the same time. This is possible if, for instance, the image contains multiple instances of the same object, in different locations. Since the saliency map represents the feedback to the feature map neurons, it is important to avoid that multiple regions be selected within a single time step, otherwise different regions of the image would become synchronized. A simple solution to this problem is to define the weights $w_k$ in eq. 13 as a function of the overall sum of the intensity values of the conspicuity maps:

$$w_k = s(\sum_{u,v} C^k_{u,v}). \tag{14}$$

The function $s(\xi)$ may for instance be a sigmoidal function such as $1 - \frac{1}{1+\exp -(\xi - \eta_w)}$, which rapidly approaches 0 as the value of $\xi$ exceeds a threshold $\eta_w$.[4]

To conclude, the model proposed for the feature maps neurons in section 5 allows a considerable simplification of the tasks performed by the conspicuity maps and makes a slow, iterative integration process no longer necessary for the construction of the saliency map. For this reason, the neuronal model for the conspicuity maps will simply be some output function of their "intensity" value, computed through eqs. 10 and 12. Similarly, the model for the saliency map neurons will be some instantaneous transfer function of the intensities $S_{x,y}$ defined by eqs. 13, 14. This procedure would allow to select target objects

[4] The threshold $\eta_w$ can be defined as a fixed percentage of the input image size.

for most synthetic images used in typical psychophysical experiments. However it is possible that, especially for input images containing noisy, highly textured patterns, more complex schemes may be needed for the integration of the conspicuity maps. In these cases, some adapted version of the relaxation process introduced in [22] may be implemented, by means of lateral intra–map and inter–map connections between conspicuity map neurons. The advantage, with respect to the instantaneous, synchronous framework of [22] is that one or two iterations of this recursive connection scheme would be sufficient to reduce noise and to select the target region. We suggest that the strengths of the intra–map connections may define some center–surround receptive fields, with a very wide negative lobe. The inter–map connections for neurons $C^k_{x,y}$ may simply be limited to excitatory links from/to all neurons $C^h_{x,y}$, $\forall h \neq k$ at the same location.

# 9 Feedback

The role of the neurons in the saliency map is to select a relevant region in the image and to increase the temporal correlation of all feature map neurons located in such region. This is done by modifying the parameters of the probabilistic model that determines their firing time, in a proportional amount to the activity level of the saliency map neurons. In particular, high activity in a neuron $S_{x,y}$ should be used as a reinforcement signal that allows to decrease all "uncertainty" widths $\sigma^k_{x,y}$, $k = 1, ..., K$. This can be done through the following updating rule:

$$\tau \frac{d\sigma^k_{x,y}(t)}{dt} = -(\sigma^k_{x,y} - \sigma_{min}) \cdot s(S_{x,y}) + \lambda_\sigma(\sigma_{max} - \sigma^k_{x,y}). \tag{15}$$

The first term represents the feedback that affects all locations at which the saliency map has a high value. The values of $S_{x,y}$ are mapped into $[0, 1]$ through a squashing function such as $s(\xi) = 1/(1 + \exp -(\xi - \eta_\sigma))$, whose value is about zero everywhere, except for values of saliency that exceed a threshold $\eta_\sigma$, for which it is $\approx 1$.

The second term in equation 15 is a positive increase of $\sigma^k_{x,y}$ that affects all feature map neurons, at a rate proportional to $\lambda_\sigma$. The consequence of this term is that, unless a location $x, y$ is selected by the saliency map, all feature map neurons at that location will gradually lose their oscillatory property, making the probability models $p^k_{x,y}$ of their firing times gradually approach a uniform distribution. However, if the $x, y$ location is at some time $t$ selected by the saliency map, the strong decrease in $\sigma^k_{x,y}, k = 1, ..., K$ will increase the periodic character of the corresponding distributions, making the corresponding neurons fire even more precisely at the next period $t + 1/\omega^k_{x,y}$.

The role of feedback with respect to the frequency parameter has been described in section 5 as the possibility to increase inter–map synchrony at a selected location $x, y$ by forcing the values of all $\omega^k_{x,y}, k = 1, ..., K$ to be integer fractions of a frequency $\overline{\omega}_{x,y}$. This driving frequency is defined as the highest frequency over a local euclidean neighborhood of radius $\rho$. More precisely:

$$\overline{\omega}_{x,y} = \max_{h,u,v} \ \omega^h_{u,v} \cdot s(S_{u,v}), \tag{16}$$

where the indices of the *max* operator vary over $h = 1, ..., K$ and $(u, v) \in \mathcal{N}(x, y) = \{(u, v), \|(u, v) - (x, y)\|_2 \leq \rho\}$. This driving frequency $\overline{\omega}_{x,y}$ is used to determine, for each $p^k_{x,y}$, $k = 1, ..., K$, the new frequency value towards which it should be shifted. The limit frequency for $\omega^k_{x,y}$ is defined as the integer fraction of $\overline{\omega}_{x,y}$ closest to $\omega^k_{x,y}$. This corresponds to the following updating rule:

$$\tau \frac{d\omega^k_{x,y}(t)}{dt} = s(S_{x,y}) \cdot (-\omega^k_{x,y}(t) + \frac{\overline{\omega}_{x,y}(t)}{n}), \tag{17}$$

where

$$n = arg \min_{m=0,1,2,...} \left| \omega^k_{x,y} - \frac{\overline{\omega}_{x,y}(t)}{m} \right|. \tag{18}$$

Since the driving frequency $\overline{\omega}_{x,y}$ is obtained over a neighborhood $\mathcal{N}(x, y)$, this rule provides some sort of local "smoothing" of the frequencies $\omega^k_{u,v}$, $k = 1, ..., K$ throughout the neighborhood.

The final parameter of the neuronal model that is affected by feedback is the phase $\phi$. Again, an updating rule that forces all phases to be coherent within a local neighborhood is most appropriate (see also the comparator model introduced in eq. 2). This can be done by forcing the phases of a neuron at location $(x, y)$ towards the phase of the neuron selected out of the neighborhood $\mathcal{N}(x, y)$ for the driving frequency in eq. 16. In other words the values of $\phi^k_{x,y}$, $k = 1, ..., K$ must approach a value $\overline{\phi}_{x,y} = \phi^\gamma_{\alpha,\beta}$, where the indices $\gamma, \alpha, \beta$ correspond to:

$$\gamma, \alpha, \beta = arg \max_{h,u,v} \ \omega^h_{u,v} \cdot s(S_{u,v}). \tag{19}$$

The updating rule on $\phi^k_{x,y}$ can thus be defined as follows:

$$\tau \frac{d\phi^k_{x,y}(t)}{dt} = s(S_{x,y}) \cdot (\phi^k_{x,y}(t) - \overline{\phi}_{x,y}(t)). \tag{20}$$

If the updating rules (15)–(20) are applied to the neurons selected by the saliency map, the synchronization of their firing patterns will considerably increase. This can be seen as an increase in their correlation coefficient, as well as in an increase of their power spectrum components around their central frequency $\omega_i$. On the other hand, neurons that are never selected by the saliency map will continuously decrease their oscillatory properties, and their firing patterns will eventually become completely decorrelated.

We conclude this section by observing that the feedback loop has also the effect of increasing the stability of the system, in terms of being capable of selecting and synchronizing regions of the feature maps with an increasing degree of accuracy. Indeed, if at an initial time a region of the image is selected thanks to the initial synchronization of some neurons within a particular feature map $k$, then the same neurons will also fire synchronously after some period of time $\Delta t$. However, the feedback also affects neurons at the same locations in other feature maps, so that, after some integer multiple of $\Delta t$, there will be even more neurons firing synchronously. This means that the same region will thereafter be selected even more accurately, since additional feature maps $h \neq k$ may contribute to its saliency (cf. eq. 13).

# 10 Conclusions and future work

In this report we have analyzed the feature binding problem, a combinatorial complexity problem that affects connectionist networks using multiple topographic representations. We have proposed that a solution to this problem may derive by the combined use of attention mechanisms and by exploiting the temporal synchrony of neuronal firing. To this end, a new framework has been proposed in in terms of a neuronal model, and of a computational architecture capable of producing synchronized firing in distributed assemblies of neurons, representing objects of interest. The system is structured into a set of feature maps, providing the input to a corresponding set of conspicuity maps, and a saliency map. In turn, the saliency map is connected to the feature maps. This feedback loop has the effect of reinforcing synchronous firing of neurons belonging to the same region of attention. With respect to similar work recently proposed in the literature, there are two main novelties in the proposed framework: the combination of neuronal oscillation/synchronization with attentive mechanisms, and the model for oscillatory feature map neurons.

Although the relationships between attention and synchronization have been formulated several years ago (see for instance [4]), very little work has been done in this direction. One interesting exception is represented by [26] [27] (cf. section 3.2), in which an external saliency map is used to enforce neuronal synchronization within a single feature map that simply reproduces the retina. Other models have been proposed, that are possibly more structured, in that they use multiple maps and complex connection schemes, such as [2]. However, the lack of inter–map connections in this system prevents from synchronizing neurons in different maps, and eventually from binding the corresponding features. Multiple maps with inter–map connections schemes are indeed proposed in several architectures based on dipoles units (e.g. [21]). However, in these cases the lack of attention mechanisms strictly limits the number of distinct phases/frequencies on which neurons may synchronize. This means that in all these models synchronization may only occur in very simplified images, for instance images contains one or two uniform objects on a uniform background.

The second novelty of the proposed framework is the neuronal model for the feature maps. Each neuron is characterized by a probabilistic description of its firing time, whose initial parameters are determined by convolution of the image with a filter bank. With respect to previous abstract models such as [2] [15], the proposed one makes explicit use of different frequencies in neuronal firings, in addition to different phases. Although the proposed stochastic model of neuronal firing is a mathematical abstraction, it corresponds well to the properties of biological neurons, and it may easily be implemented with computational neurons, for instance through dipoles. The advantage of this model with respect to dipoles is that it allows a wider range of dynamics, mainly through the value of the parameter $\sigma$, defining the "uncertainty" of each neuron. By varying this parameter, it is possible to have a soft transition from a neuron having a highly periodic, correlated firing pattern, to an aperiodic one, decorrelated from the remaining neurons.

The work presented in this report, however, is more a theoretical framework than a complete solution to the binding problem. For instance, the mechanisms proposed for each step need to be extensively evaluated by testing them on a variety of input images. The pros and cons of this strategy with respect to the instantaneous, synchronous framework of the previous attention system [22] will also have to be assessed. As far as the theoretical model

is concerned, more work will be required in order to define a rapid relaxation algorithm for the integration of the conspicuity maps. Finally, some issues of the proposed framework have not been covered at all in this study. They concern the definition of modules for the representation of feature conjunctions and for object recognition. For instance, mechanisms need to be defined for the detection of synchronous, periodic patterns of activation in feature map neurons. In this way, it may be possible to learn conjunctive descriptions of features for objects of interest, if their images have been presented to the network for a sufficient amount of time.

**Acknowledgements.** I would like to thank Lokendra Shastri for inspiring this project, and for extensive discussions on all these issues. I am also grateful to Jitendra Malik and Jerome Feldman for useful comments.

## References

[1] C.H. Anderson and D.C. Van Essen, *Shifter Circuits: A Computational Strategy for Dynamic Aspects of Visual Processing.* Proc. of National Academy of Sciences USA, Vol. 84, 1987, pp. 6297-6301.

[2] P. Baldi and R. Meir, *Computing with Arrays of Coupled Oscillators: An Application to Preattentive Texture Discrimination.* Neural Computation, Vol. 2, No. 4, 1990, pp. 458-471.

[3] T. Carney, M.A. Paradiso, and R.D. Freeman, *A Physiological Correlate of the Pulfrich Effect in Cortical Neurons of the Cat.* Vision Research, Vol. 29, No. 2, 1989, pp. 155-165.

[4] F. Crick, *Function of the Thalamic Reticular Complex: the Searchlight Hypothesis.* Proc. of the National Academy of Sciences, 1984, 81:4586-4590. Also in Anderson and Rosenfeld (eds), *Neurocomputing.* MIT Press, 1988, pp. 569-575.

[5] R. Eckhorn, R. Bauer, W. Jordan, M. Brosch, W. Kruse, M. Munk, and H.J. Reitboek, *Coherent Oscillations: a Mechanism of Feature Linking in the Visual Cortex?* Biological Cybernetics, 1988, 60:121-130.

[6] R. Eckhorn, H.J. Reitboeck, M. Arndt, and P. Dicke, *Feature Linking via Synchronization among Distributed Assemblies: Simulations of Results from Cat Visual Cortex.* Neur. Comp. Vol. 2, pp. 293-307, 1990.

[7] C.W. Eriksen and J.D. ST. James, *Visual attention within and around the field of focal attention: A zoom lens model.* Perception and Psychophysics, 1986, Vol. 40, No. 4, pp. 225-240.

[8] E.R. Grannan, D. Kleinfeld, and H. Sompolinsky, *Stimulus-Dependent Synchronization of Neural Assemblies.* Neur. Comp. Vol. 5, pp. 550-569, 1993.

[9] C.M. Gray and W. Singer, *Stimulus–Specific Neuronal Oscillations in Orientation Columns of Cat Visual Cortex.* Proc. Natl. Acad. Sci. USA, 1989, Vol. 86, pp. 1698-1702.

[10] C.M. Gray, P. Koenig, A.K. Kreiter, T.B. Schillen, and W. Singer, *Oscillatory Responses in Cat Visual Cortex Exhibit Inter–Columnar Synnchronization which Reflects Global Stimulus Properties.* Nature, 1989, Vol. 338, pp. 334-337.

[11] S. Grossberg and D. Somers, *Synchronized Oscillations During Cooperative Feature Linking in a Cortical Model of Visual Perception.* Neural Networks, 1991, Vol. 4, pp. 453-466.

[12] D. Horn, D. Sagi, and M. Usher, *Segmentation, Binding and Illusory Conjunctions.* Neural Computation, 1991, Vol. 3, No. 4, pp. 510-525.

[13] J.E. Hummel and I. Biederman, *Dynamic Binding in a Neural Network for Shape Recognition.* Psychological Review, Vol. 99. No. 3, 1992, pp. 480-517.

[14] B. Julesz, *Towards an axiomatic theory of preattentive vision.* In Edelman et *al.* (eds), *Dynamic Aspects of Neocortical Function.* Wiley, 1984, pp. 585-612.

[15] D.M. Kammen, P.J. Holmes, and C. Koch, *Cortical Architecture and Oscillations in Neuronal Networks: Feedback Versus Local Coupling.* In: R.M.J. Cotterill (edt.), *Models of brain function.* Cambridge University Press, 1989.

[16] C. Koch, *Computational Approaches to Cognition: The Bottom-Up View.* Curr. Opin. Neurobiol. 3:203-208, 1993.

[17] C. Koch and H. Schuster, *A Simple Network Showing Burst Synchronization without Frequency-Locking.* Neur. Comp., Vol. 4, pp.211-233, 1992.

[18] J.M. Kowalski, G.L. Albert, B.K. Rhoades, and G.W. Gross, *Neuronal Networks with Spontaneous, Correlated Bursting Activity: Theory and Simulations.* Neural Networks, Vol. 5, 1992, pp. 805-822.

[19] J. Malik, and P. Perona, *Preattentive Texture Discrimination with Early Vision Mechanisms.* Journal of Optical Society of America, Vol. 7, No. 5, 1990, pp. 923-932.

[20] C. von der Malsburg and W. Schneider, *A Neural Cocktail–Party Processor.* Biological Cybernetics, Vol. 54, 1986, pp. 29-40.

[21] C. von der Malsburg and J. Buhmann, *Sensory Segmentation with Coupled Neural Oscillators.* Biological Cybernetics, Vol. 67, pp. 233-242, 1992.

[22] R. Milanese, *Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation.* Ph.D. thesis, Univ. of Geneva, 1993. (Anonim. ftp: cui.unige.ch, pub/milanese/thesis).

[23] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost and T. Pun, *Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation.* Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, June 20-24, 1994. Also available as Technical Report TR-94-14, International Computer Science Institute, March 1994.

[24] J. Moran and R. Desimone, *Selective Attention Gates Visual Processing in the Extrastriate Cortex.* Science 229, 1985, pp. 782-784.

[25] M.C. Mozer, *The Perception of Multiple Objects: a Connectionist Approach.* MIT Press, Cambridge, MA, 1991.

[26] E. Niebur, C. Koch, and C. Rosin, *An Oscillation–Based Model for the Neuronal Basis of Attention.* Vision Research, Vol. 33, No. 18, 1993, pp. 2789-2802.

[27] E. Niebur and C. Koch, *A Model for the Neuronal Implementation of Selective Visual Attention Based on Temporal Correlation Among Neurons.* Journal of Computational Neuroscience, to appear.

[28] H. G. Schuster (edt.), *Non-Linear Dynamics and Neuronal Networks.* VCH, Weinheim, 1991.

[29] L. Shastri and V. Ajjanagadde, *From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables and Dynamic Binding Using Temporal Synchrony.* Behavioral and Brain Sciences, 1993, Vol. 16, pp. 417-494.

[30] W. Singer, *Synchronization of Cortical Activity and Its Putative Role in Information Processing.* Ann. Rev. Physiol. 55:349-374, 1993.

[31] S.J. Thorpe, *Feed–Forward Processing in the Visual System and the Role of Asynchrony.* Symp. on Dynamics of Neural Processing, Washington D.C., June 6-8, 1994, pp. 151-155

[32] A.M. Treisman and G. Gelade, *A Feature–Integration Theory of Attention.* Cognitive Psychology, 12, 1980, pp. 97-136

[33] J.K. Tsotsos, *Analyzing Vision at the Complexity Level.* Behavioral and Brain Sciences, Vol. 13, 1990, pp. 423-469.

[34] D.L. Wang, J. Buhmann, and C. von der Malsburg, *Pattern Segmentation in Associative Memory.* Neur. Comp., Vol. 2, pp. 94-106, 1990.

[35] J.M. Williams and A. Lit, *Luminance–Dependent Visual Latency for the Hess Effect, the Pulfrich Effect, and Simple Reaction Time.* Vision Research, Vol. 23, No. 2, 1983, pp. 171-179.

[36] J.M. Wolfe, K.R. Cave and S.L. Franzel, *Guided Search: An Alternative to the Feature Integration Model for Visual Search.* Journal of Experimental Psychology: Human Perception and Performance, Vol. 15, No. 3, 1989, pp. 419-433.