Attention

Visual Attention acts to select certain types of features and certain locations for enhancement. This process is thought to speed up and improve the quality of object recognition processes by suppressing irrelevant information and enhancing relevant information concerning the nature of an object.

What determines which features and spatial locations are to be attended to?

- Exogenous cues (image driven, popout)
- Endogenous cues (priming, volition)

Exogenous Attention

In exogenous attention, the focus of attention is determined by what is in the retinal image, or camera signal.

Humans are extremely efficient in detecting a part of an image that differs in a single aspect from the rest of the image, implying parallel processing.

This phenomenon is called $popout$ or "preattentive vision".

For example, a green letter "P" in a field of red letter "T"'s is detected quickly, with the detection time being independent of the number of distracting elements.

Conjunctions of features are more difficult to detect, and require a serial search through all possible objects, sometimes referred to as "attentive vision".

For example, detecting a red letter "P" among a field of red and green letter "T"s is difficult, with the detection time linearly increasing with the number of distractors.

These phenonmena imply that regions of the image that differ in a single feature value from the rest of the image "attracts attention".

Endogenous Attention

In endogenous attention, the focus of attention is determined by processes within the brain, such as memory or task planning.

Experiments have shown that some neuron‘s responses are attenuated according to the similarity of the current stimulus to memory traces. This could be used to draw attention to novel stimuli.

Priming is an attentional effect whereby repeated stimuli at a location primes shifting of attention to that location in the future.

People are also able to volitionally shift their attention to a given location or to attend to a given type of feature (such as color or size).

Attentional Modulation in Neurons

(Moran, J. and Desimone, R, (1985), "Selective attention gates visual processing in the extrastriate cortex", *Science*, Vol. 229, pp 782-784)

Many neurons in the higher levels of the visual cortex have very large receptive fields (i.e. they respond to input over a large area of the image).

This implies that a single such neuron would produce a rather ambiguous response when stimulated by a scene containing a lot of objects.

Moran and Desimone found that some neurons in cortical areas V4 and IT had responses to multiple stimuli within their receptive fields that depended on their allocation of attention.

When two stimuli were present in the receptive field, the neuron's response was determined primarily by the attended stimulus, while the influence of the unattended stimulus was greatly attenuated.

Example: Suppose a neuron was selective to the colour red, and there was a red stimulus and a green stimulus within its receptive field. If attention was directed (say by volition) to the location of the red stimulus, then the neuron would respond strongly. If attention was directed to the location of the green stimulus then the neuron would respond weakly.

Note that the stimulus has not changed at all, only the location of attention.

This attentional effect only occurs when the stimuli are both within the receptive field of the neuron. If there is only one stimuli in the RF then attention has no effect.

Saliency Map Models of Attention

(Koch, C. and Ullman, S. (1985), "Selecting one among the many: A simple network implementing shifts in selective visual attention", MIT AI Lab Memo 770, Jan. 1984)

In the Koch and Ullman model of attention, selective attention is used to enhance the response of a set of feature detector neurons at a certain location that are passed to a location insensitive "Central Representation".

In this model, attention is directed to the most "Salient" image location.

The term *Saliency* refers to the ability of an image location or image feature to attract attention.

For example, a red blob surrounded by red blobs would be less salient that a red blob surrounded by green blobs.

A *Saliency Map* is a topographic map which has, at each location in the map, a measure of the saliency at that location.

The *most* salient location can be found by finding the maximum saliency value in the saliency map. In the Koch and Ullman model this is done with a Winner-Take-All process.

Winner-Take-All

A Winner-Take-All process is one which maps a set of input values into an set of output values (of equal number) using the transformation rule:

$$\begin{array}{ll} y_i = 0 & x_i < \max_j\{x_j\} \\ y_i = f(x_i) & x_i = \max_j\{x_j\} \end{array}$$

where $\{x_i\}$ is the input set, $\{y_i\}$ is the output set, and $f(x)$ is some constant or increasing function.

The Koch and Ullman WTA involves two binary pyramid structures. Each pair of elements in the first pyramid are compared and the maximum value sent to the next level in the pyramid. The global maximum is propagated in this way to the top of the pyramid. The flow of information in the second pyramid is downward, as the "path" of the most active elements in the first pyramid is marked, terminating at the "winning" element at the lowest level.

Linear Network WTA Implementations

Koch and Ullman introduce a linear (as opposed to their pyramidal structure) network for implementing the WTA.

$$x_i^{t+1} = f\left(I_i^t + \sum_{j \neq i} w_{ij} x_j^t\right)$$

where x_i^t is the value of the WTA output at location i at time t, w_{ij} is the "synaptic weight" between the output of the unit at location j and the input to the unit at location i, and I_i^t is the feature value at location i at time t.

Unfortunately, this network is unstable and may oscillate.

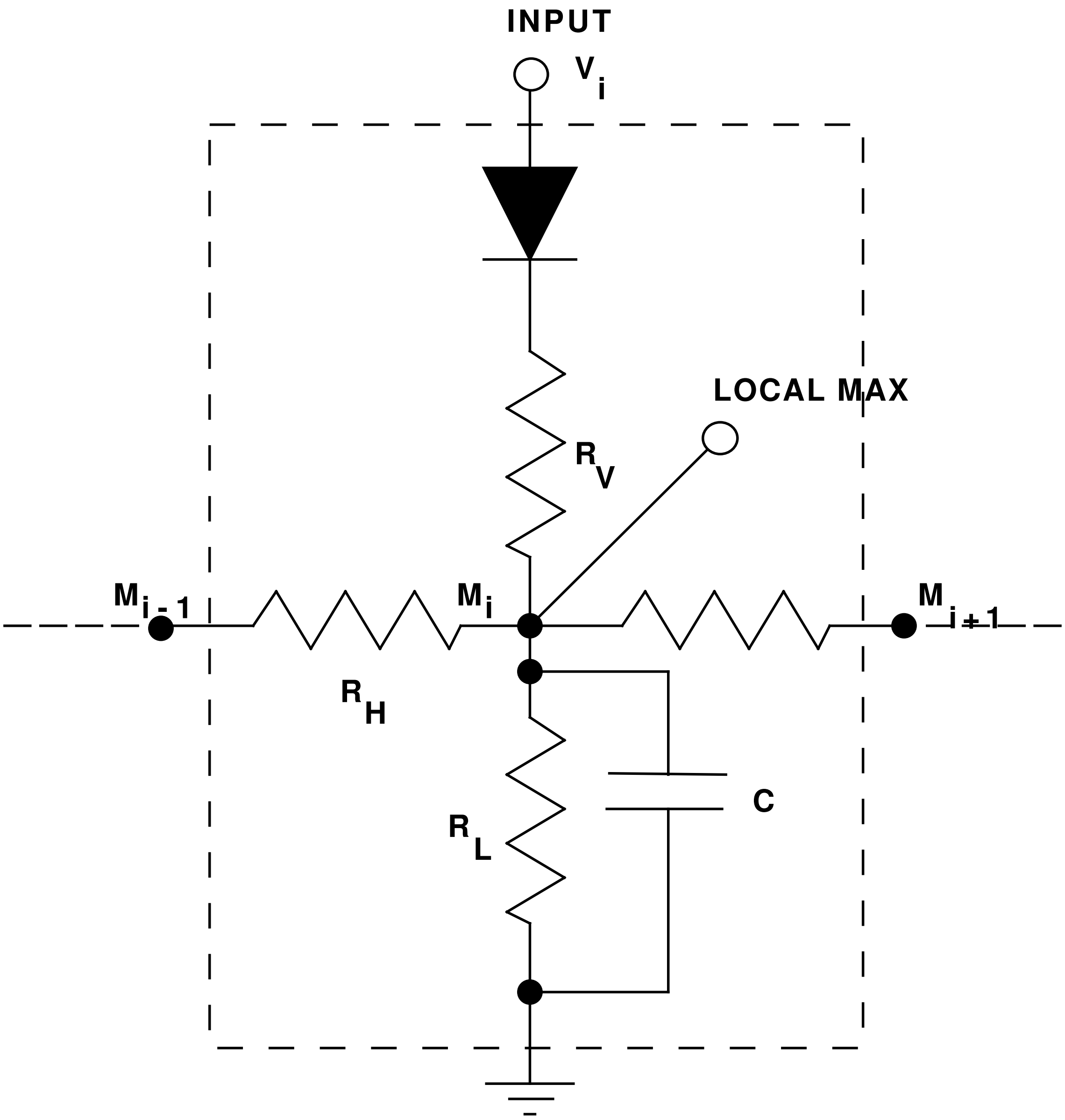
Another network approach, which IS stable, is the following:

$$x_i^{t+1} = x_i^t + K * \mathsf{S}\left((\alpha \max(x_i^t * I_i^t) - x_i^t * I_i^t)\right)$$

where $\mathcal{S}$ is a saturation nonlinearity which prevents x from going above and below upper and lower limits. The parameter α is a constant that is in the range (0,1).

If the value of $x_i * I_i$ is less than the maximum value times α then x_i will decrease. If $x_i * I_i$ is equal to the maximum value then x_i will increase. This results in a form of positive feedback, leading to Winner-Take-All behaviour.

The maximum operation can be implemented by the network shown in the figure.



The Maximum Network