Visual Motion Detection Based on a Cooperative Neural Network Architecture

Robert PALLBO
Dept. of Computer Science, Lund University, Box 118, S–221 00 LUND, Sweden
E-mail: robert.pallbo@dna.lth.se

Abstract. A neural network architecture for visual direction detection is proposed. The approach assumes a continuous flow of visual stimuli as input. The output of the network will as a consequence be a continuous flow as well. Each node in the network signals when motion is occurring at their position. Other nodes make use of this information in their computations. This allows, from a computationally viewpoint, rather simple nodes. When a node detects a motion, this detection is propagated through the network in the direction of the motion. Such a propagation is easy to construct. The initial detection of a motion is carried out by spontaneous activity among the nodes. Hence, no motion detection is carried out by the nodes, but are an emergent property of the collaboration in the network. In this paper, the model is presented and results from a computer simulation of the process is discussed. Related models of direction selectivity are also discussed in relation to the proposed model.

1 Introduction

The field of computer vision comprehends many hard problems. One of them is the detection of moving objects in the observed environment. Several approaches to this problem have been suggested. Some of those are using techniques from the area of neural networks and, among these, there are models closely related to models of biological vision. The model presented in this paper belongs to this category.

One of the virtues of neural networks is parallel computation. Each node in the network have access only to a fraction of the total amount of the input data. Such an arrangement is especially useful in computer vision and is easily achieved because of the distributed nature of image pixels. In this paper, the input will consist of a continuous flow of successive images. As a consequent, the output of the computation will share this property. This feature will be central to the approach taken in this paper.

The motion to be detected will be assumed to arise from objects in the observed environment. If one part of an object is moving in one direction, the other parts must do as well. Assuming the objects to have a significant size relative to the visual field, the following observation can be made: *A picture part can be moving in a certain direction only if some adjacent picture parts are moving in the same direction.* This observation will be considered in what follows.

The model to be presented, is to a great extent inspired by biological models and studies of visual cortex. These are especially the early work on rabbit's retina (Barlow and Levick, [2]), the classical work on simple and complex cells (Hubel and Wiesel, [10]), the dynamics

of visual cortex (Gilbert et al., [8]), studies of the spontaneous activity in cortical neurons (Evarts, [5]), the cross-inhibitory model of orientation selective cells (Ferster and Koch, [6]) and the theory of neural Darwinism (Reeke Jr. et al., [14]). Other influences have been studies of army ants (Franks, [7]) and Braitenberg's delightful book on vehicles [4].

2 Terminology

For the matter of simplicity in the discussion, this section will introduce some useful terms. The networks that will be proposed in this paper are arranged in two-dimensional layers. We will primarily make use of two such layers, one for the visual input, and one resultant layer for the output. Each node in the network receives input from a lot of other nodes. These can be separated into sets depending on how they will affect the target node. These sets will be called *neighbourhoods* to indicate their topological arrangement (cf. Cellular automata, and Kohonen networks, [11]). For the purposes of this paper, two such neighbourhoods are defined:

$N_T(i)$ is the set of nodes in the *input* layer that are connected with the target node i. (T denotes topological.)

$N_L(i)$ is the set of lateral connected nodes, i. e., the nodes in the *output* layer connected with the target node i.

An additional index $+$ and $-$ will denote the subsets of a neighbourhood with excitatory respectively inhibitory influence on the target node i. In the following $N_{L+}(i)$ and $N_{L-}(i)$ will be used.

Each node acts basically as a simple threshold device with spontaneous activity as an additional feature. We will differentiate between the internal potential of the node, and the (external) output. The internal potential of a node i, μ_i, is computed at every iteration by:

$$\mu_i = \eta \cdot \mu_i + \sum_k (w_{ki} \cdot x_k) \tag{1}$$

where w_{ki} is the weight of the connection from node k to node i, and x_k is the output of node k. The external output of the node, x_i, is computed by:

$$x_i = \begin{cases} 1 & (\mu_i \geq \theta) \text{ or } (P_i(t) = 1) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where θ is a threshold value and $P_i(t)$ is a stochastic variable determining the spontaneous activity of the node.

Each of the nodes within a neighbourhood, will be connected with the same weight to the target node. These weights will be different for various applications. The weights can be expressed with constants $\omega^{(X)}$, where X corresponds to the index of the neighbourhood $N_X(i)$ in which the node source node residents. More formally:

$$w_{ki} = \begin{cases} \omega^{(X)} & k \in N_X(i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

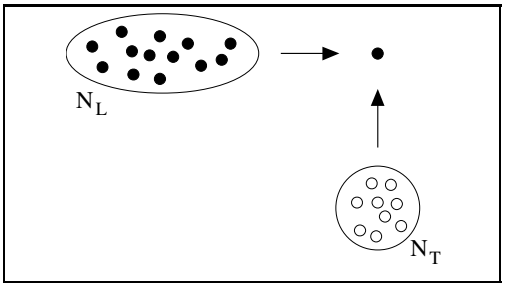

Figure 1: The principal topology of the proposed model. Two layers are considered, the input layer and the output layer of nodes selective to one specific direction. The black circles represent output nodes and the white circles represent input nodes. The stand alone output node represents the target node. The picture shows two neighbourhoods, a lateral neighbourhood (N_L) and a topological (N_T). The nodes in N_L are all situated left of the target node causing the node to be selective to motions from the left to the right. The nodes in N_T are all situated below the target cell.

3 A Proposal for a Cooperative Model of Motion Detection

The sets defined above, will now prove valuable as we describe an architecture for detecting visual motion. The nodes in the model to be proposed, does not detect motion in an arbitrary direction, but are specialized to detect different directions. Hence, the output layer will actually be composed of several layers, one for each direction.

Usually in computer vision, the input layer is analyzed and compared to a reference. In the model proposed here, the output layer will be used in the analyzing. The network is constructed out of nodes that in isolation would not be able to act as direction or motion selective units. When gathered and cooperating by means of lateral connections, such detection is an emergent property of the network.

As in any model of vision, the neighbourhoods of the nodes in the model must, at least, include one topological set that makes out the input. The topological shape of this set, as well as the usage, differs in various models. Here, the set, N_T, will be radial symmetric and centered to the topological position of the target node (Fig 1). Therefore, only a plain projection of the input will take place. No features will derive from this arrangement.

Besides the projection of input, a lateral set, N_L, is connected. Only nodes in the output layer are included in this set. The set will be asymmetrical and positioned directly beside the topological position of the target node. The side of the target node that the set resides in, determines the direction of motion that the target node will respond to. In the set N_L, only nodes that responds to the same direction as the target cell are included. This is crucial to the function as will be shown later.

This completes the architecture. Only these two excitatory sets are needed. The architecture can be interpreted as follows: A motion in the visual field is not stationary per definition. It is also the case that the output layer is topological preservative. Therefore, the indication of a motion must track the object as it propagates in the visual input layer. The main purpose for the node in the output layer, is to propagate this detection further. To accomplish this, each node simply detect simultaneous activity in N_L and N_T. When such simultaneous activity occurs, the node will activate.

Once a motion has been detected, it is clear that the architecture will ensure the further detection. The initial detection, on the other hand, must be dealt with otherwise. This is where the spontaneous activity is found useful. In equation 2 we used a stochastic variable P that introduces "noise" into the system. This added activity ignites the detection, and once ignited, it will spread. In the case of a false ignition, the chain reaction will not continue because of inappropriate stimuli at the input layer.

To visualize the model, the architecture has been tested in a computer simulation. The two layers were both matrixes of size $10 * 100$ pixels. As stimulus, a bar of $10 * 2$ pixels was used. This bar was either propagating left, right or being stationary. The output layer included only nodes selective to movements to the right, no other direction was implemented. This was possible since no inhibitory connection from nodes tuned to other directions are necessary. The topological set, N_T, consisted of one node in the input layer. This node was connected with a strength $\omega^{(T)} = 900$. The lateral set, N_L, consisted of a $4 * 10$ matrix of nodes directly to the left of the target node. These connections were given the strength $\omega^{(L)} = 50$. In addition, the threshold $\theta = 1000$ and the probability was 1.0% that $P_i(t) = 1$.

3.1 Results

The above arrangement worked well. The simulator environment allowed for an investigatement of what features that are crucial for the performance of the network. These features were found to be the width of the bar, the speed of the bar, the shape of N_L and the arousal level of the spontaneous activity.

The default width of the bar was set to two pixels, or nodes, as mentioned above. There were no significant effects when decreasing the width to one pixel. When increased, on the other hand, a critical level was found. This level is determined by several factors. When the bar covers most of N_L, then the lateral input gets too strong and the output nodes can be activated by this stimulus alone. If so, the activity will spread to other nodes, regardless of the activity in the input layer, and something like an epileptic seizure will occur. A rule of the thumb is that the bar should be only a fraction of the width of the N_L sets. To avoid this problem associated with wide bars, one must filter the input image. This filter should be selected so that only zero-crossings, or edges, will appear in the filtered image. In the mammal visual system, such a filtering are made by the ganglion cells situated on the retina. The activity in these cells are projected to the visual cortex where, among others, motion detecting cells are resident.

If the speed of the bar is too low, the network consider the bar as stationary. The net requires that the bar propagates at least one pixel at every iteration. In the other extreme, if the bar moves too fast, the net will not be able to detect the movement. The network will simply not consider the activity to be caused by the same object. Fast movement in the simulation is when the bar does not propagate from pixel to pixel, but skips several pixels at each iteration. The critical limit is reached when the bar skips more pixels than the lateral sets are wide. A bar must appear at least once in every lateral set, otherwise the network is unable to propagate the detections in the output layer. The ability to detect fast movement is thus closely related to the topology of the lateral sets. If these sets are too small, the stimuli must propagate slowly in order to appear in every set. If the lateral neighbourhoods grow too large, however, the resolution in the system will suffer. One has thus to make a compromise between resolution and speed. Another possibility would be to use more than one output layer for each direction to be detected. It would be possible to use one layer with high resolution but slow speed and

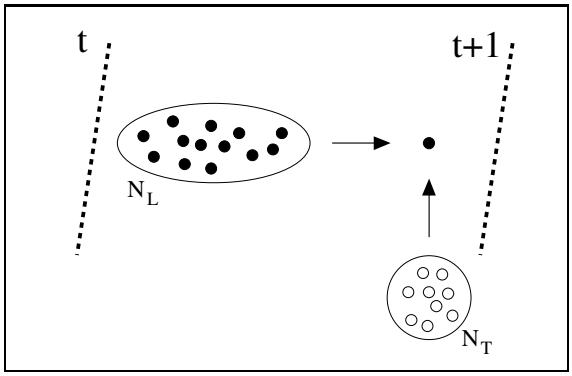

Figure 2: Illustration of the problem associated with fast moving stimuli. The position of the stimulus are shown at time t and $t + 1$. No nodes can mediate the detection from the stimulus position at time t to the position at time $t + 1$.

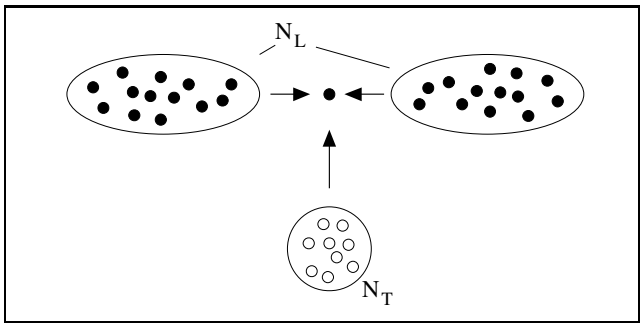

Figure 3: The principal topology of the network when modified for detection of edge orientation. Two lateral neighbourhoods are used instead of one as in the case of direction detection.

another layer with low resolution but high speed. These layers would then complement each other.

It is obvious that the level of spontaneous activity cannot be too large. In that case, there would simply be too much noise in the system and the signal/noise ratio be too low. When the activity is too low, on the other hand, it will take longer before any initial detection of a movement will occur. Therefore, the level of the spontaneous activity can serve as a filter of motion. With a low level, motion that appears only for a short period will not be detected. With a high level, it will. What level to choose is dependent on the specific application in which the system is used.

The network architecture can be modified to compute the orientation of edges instead of motion. This is achieved by changing the topology of N_L to reside on both sides of the target node. In this arrangement it is necessary to introduce cross-inhibition, i. e., inhibitory connections between nodes of different orientational tuning. The basic operation is the same as in the motion detection arrangement. When this model was tested in a computer simulation, temporal summation was used since this gave a better performance [13].

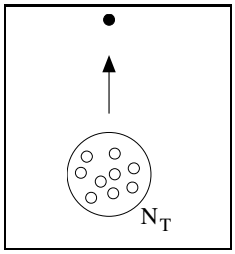

Figure 4: The principal topology of gradient schemes. No lateral connection is used. The topological connection is used in a different way than the model proposed in this paper. Refer to the text for details.

4 Related Models of Motion Detection

In the literature of computer vision, many models of motion detection are found. Those that also are claimed as biological models can be divided into two categories, the gradient and the correlation schemes. These will be discussed in relation to the model proposed in this paper. For a more detailed description of those models, refer to [16, 3, 12, 2].

4.1 Gradient Schemes

In the gradient scheme, the spatial intensity is correlated with the temporal change of intensity. The idea is that an object has a spatial pattern of intensity. When the object moves, this pattern will be reflected in the temporal change of the potential of a node. It is therefore possible to correlate this pattern with the spatial pattern, and consequently know in what direction the object is moving. One cannot only detect in what direction the object is moving, but also the speed of the motion. This is accomplished by correlating the scale of the temporal and spatial patterns.

The neural implementation of this scheme requires a rather precise architecture. Further, it does not assume any lateral connection among the direction selective nodes. If such lateral connection would be introduced, however, a less precise architecture would be possible. The nodes would get a weak indication of motion from its topological input and this indication would then be strengthened by the lateral connection. The model proposed in this paper, however, suggests that no such topological indication of motion is required in order to carry out the detection.

4.2 Correlation Schemes

The mechanism of correlation schemes includes a broad range of models. Yet, they have important characteristics in common. The principal idea is to use two distinct spatial areas in the input image and to correlate the sub-images in those. One image is recorded with a slight delay in relation to the other. If the images of the two fields correlate, it is due to motion taking place in the observed visual area. In the terms of this paper there are two topological neighbourhoods, N_{T1} and N_{T2}, but no lateral neighbourhood.

A simple example of a neural network implementation of this mechanism is when the integrated activity in N_{T1} and N_{T2} is used as input to the target node. One of these neighbourhoods are delayed one iteration. If the target node receives input from both sets simultaneously, the

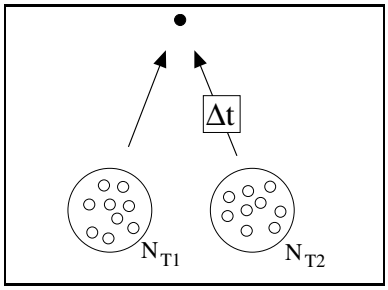

Figure 5: The principal topology of correlation schemes. One set of input nodes is delayed.

target node will activate and signal that a motion has occurred. Since one input channel is delayed, such an occurrence suggests that a motion from the delayed to the non-delayed input set has taken place. This can be expressed as,

$$\mu_i = \sum_{j \in N_{T1}(i)} \omega^{(T1)} x_j(t) + \sum_{j \in N_{T2}(i)} \omega^{(T2)} x_j(t-1) \tag{4}$$

An inhibitory variant of this scheme is also frequent in the literature. In this case, one set of connections is used to cancel the inputs from the other set. If the activity first occurs at the cancelling set, then no activation of the target node will take place. On the contrary, when the activity reaches the non-cancelling set first, the input is not cancelled, and the target node will activate and signal a motion. Hence, this mechanism is direction selective. The preferred direction of the motion is determined by the topology of the connections as indicated above.

In contrast to the model proposed in the earlier section, the correlation scheme needs input from two adjacent images. In a neural network implementation, this necessitates a usage of intermediate nodes. As a result, the architecture gets more complex. From a computational point of view, the arrangement twofolds the number of nodes. Another thing is that the correlation and gradient schemes have no interaction between the nodes in the output layer. In the proposed model, surrounding nodes can mediate a strong hypothesis that a motion is occurring in the surroundings of the target node. This can then affect the activity of the target node in the case when it only has weak indications of motion from the neighbourhood in the input layer.

5 Discussion

One might object that the spontaneous activity used in the suggested model induces a lot of false indications of movement. This is true only if a single node will have a significant effect on the further processing in the network. The model presented in this paper proposes that no such importance should be made upon one single node. Only when several nodes synchronously indicate motion, should they have any considerably effect to the process. In the proposed network, is it not sufficient that only one node in the lateral set is active, several active nodes are required to affect the target node. In the default setting of the simulation, two nodes were required active in this set simultaneously as the topological set (node) indicated activity. Simply put, several nodes must cooperate to indicate a movement.

An important property of the proposed model is the simplicity of the local calculation. Simplicity is to strive for in all scientific modelling. However, the simplicity should be within the entities, not in the behaviour [9]. The correlation mechanism might appear as simpler than the proposed mechanism, but the architectural complexity is much greater in that model.

In neural networks, a relaxation scheme is sometime used. The global energy of the system is minimized and, with some tricks (e. g. symmetrical connections), each node will know exactly how much it contributes to the global energy [1]. In the model of this paper, no such consideration of global energy has been considered. Neither do the nodes in this network have any possibility to know how much it would contribute to the global energy level. In simulated annealing, the temperature is initially set to a high degree and is subsequently decreased until a stable configuration is achieved. In the model proposed in this paper, the temperature is kept at a constant level. If the spontaneous activity is removed sometime during the process, the present motions will still be propagated until the object vanishes or halts. No new detections will occur, however, with the ignition mechanism absent.

In applications of neural network, data is distributed. This is considered as one of the greatest virtues of such network. In most such applications, though, the computation is distributed only to a limited degree. Each node is a complete machinery capable of computing whatever it is set to do. In the nodes in the network of the model presented here, this is not so. These nodes are themselves not capable of computing whether motion has occurred or not. They must cooperate with other nodes to be able to meet their obligations as motion detectors. All nodes involved in the cooperation will have the same limitation. The selectivity is therefore, a truly emergent property due to the cooperation. If any of those nodes were to be studied from outside, the conclusion most probably to be drawn would be that these nodes computes motion. They do, and do not. The solution to this paradox is a question on the level of observation of the network.

In Smolensky's viewpoint, units in connectionist models are not at the same level as neurons. Neurons are at the neural level, while the units are at the subsymbolic level. The model presented in this paper is not a connectionist model. Rather is it a model at the neural level. The behaviour of the network can anyhow be observed from a subsymbolic level, but the model is implemented at the neural level. The subsymbolic level emerges from the construction of the network, just like Smolensky argues that the symbolic level emerges from the subsymbolic level [15]. I believe that the analysis of neural networks has a lot to gain from a multi-level observation of the process. Learning, behaviour and implementation, of the very same network, should not necessary be described at the same level. The model suggested in this paper serves as an example of this. At the level of implementation, the proposed network does not compute motion, but at a higher (subsymbolic) level, this is what the nodes are signalling.

Aknowlegement

I would like to thank Lucia Vaina for her helpful direction to some relevant references.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] H. B. Barlow and W. R. Levick. The mechanism of directionally selective units in rabbit's retina. *Journal of Physiology*, 178:477–504, 1965.

[3] A. Borst and M. Egelhaaf. Principles of visual motion detection. *Trends in neuroscience*, 12:297–306, 1989.

[4] V. Braitenberg. *Vehicles, experiments in in synthetic psychology*. The MIT Press, 1984.

[5] E. V. Evarts. Temporal patterns of discharge of pyramidal tract neurons during sleep and waking in the monkey. *Journal of Neurophysiology*, 27:152–171, 1964.

[6] D. Ferster and C. Koch. Neuronal connections underlying orientation selectivity in cat visual cortex. *Trends in Neuro Science*, 10:487–492, 1987.

[7] N. R. Franks. Army ants: A collective intelligence. *American Scientist*, 77:138–145, 1989.

[8] C. D. Gilbert, J. A. Hirsch, and T. N. Wiesel. Lateral interactions in visual cortex. *Cold Spring Harbor Symposia on Quantitative Biology*, 50:663–677, 1990.

[9] P. Hogeweg. Simplicity and complexity in mirror universes. *Biosystems*, 23:231–246, 1989.

[10] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

[11] T. Kohonen. *Self-organization and associative memory*. Springer-Verlag, 1989.

[12] D. Marr. *Vision*. W. H. Freeman and Company, 1982.

[13] R. Pallbo. Neuronal selectivity without intermediate cells. *Lund University Cognitive Studies*, 13, 1992.

[14] G. N. Reeke Jr., O. Sporns, and G. M. Edelman. Synthetic neural modelling: Comparisons of population and connectionist approaches. *Connectionism in Perspective*, pages 113–139, 1989.

[15] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23, 1988.

[16] S. Ullman. The measurement of visual motion: Computational considerations and some neurophysiological implications. *Trends in Neuroscience*, 6:177–179, 1983.