A visual system for invariant recognition in animated image sequences

Luis F. Lago-Fernández [a,*], Manuel A. Sánchez-Montañés [a], Eduardo Sánchez [b]

[a] *E.T.S. de Informática, Universidad Autónoma de Madrid, Madrid 28049, Spain*

[b] *Facultad de Ciencias Físicas, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain*

Abstract

The problem of invariant recognition in animated image sequences is investigated. Some issues on this regard are discussed and a solution in terms of biological principles is proposed. It integrates a bottom-up attention module, a synchronization-based segmentation mechanism, a normalization network and a recognition module. The results show how the system provides the following properties: (1) change-based attention in animated sequences, (2) multiple object segmentation, and (3) translation and scale invariant recognition.

Key words: Visual system, Attention, Synchronization, Invariant recognition

1 Introduction

There are two basic approaches to the problem of invariant recognition: normalization (8; 12), and invariant feature extraction (1; 10). In this work we propose a normalization-based recognition mechanism that tries to preserve all the advantages of these models, while minimizing the complexity of the required network. Additionally, we have addressed how invariant recognition can be achieved when animated image sequences are presented. This situation requires from the system: (1) to determine the relevant information over time, and (2) to adapt the invariant mechanisms to work with dynamic inputs. In our approach, the relevant information is selected by a bottom-up attention

[*] Corresponding author
Email address: Luis.Lago@ii.uam.es (Luis F. Lago-Fernández).

module, which consists of a change-based saliency map and a Winner Takes All (WTA) mechanism inspired by (4). The dynamic inputs provided by the attention module are processed by a synchronization-based segmentation network (5). Segmented objects are normalized by a simple recurrent network to achieve position, scale and orientation invariant representations, which are then fed into the recognition module. Figure 1 shows a general overview of the system. In the following section we discuss the four main modules.

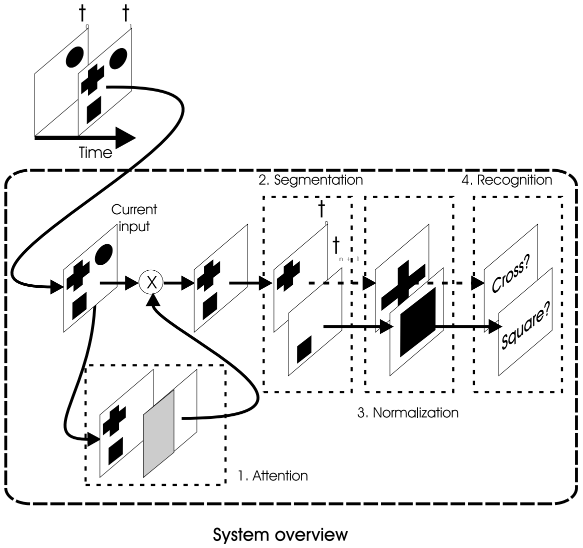

Fig. 1. System overview. Four main parts can be distinguished: (1) change-based attention module, which eliminates information about static regions; (2) segmentation module, which separates the different objects in the scene; (3) normalization module, which provides position, scale and orientation invariant object representations; and (4) recognition module.

2 The system

In our approach a change-based feature map is proposed to deal with dynamic images and detect changing regions. Figure 2 shows how these regions drive a WTA process which facilitates relevant locations for further detailed processing. The attention module has been adapted from (3) to deal with changing images. The temporal derivative of the input image $\dot{x}_{ij}$, is computed by means of a set of interneurons that store the previous image and provide, at any time t, an inhibitory input to the neurons in the saliency map. The output d_{ij} of the change-based saliency map is then defined by $\dot{x}_{ij}$, being active ($d_{ij} = 1$) if this derivative is greater than a threshold Θ, and silent ($d_{ij} = 0$) otherwise. In the WTA map, each control neuron attends a predefined region of the saliency map. The control neurons compete among themselves through a global feedback inhibition (see figure 2), and have recurrent connections (not shown in the figure) to recall the previous winner locations. These connections are disabled when afferent activity is detected. Finally, the connectivity between the input neurons and the output neurons is controled by a presynaptic facilitation mechanism triggered by the WTA map. Thus the connection between x_{ij} and o_{ij} is only activated when the activity of the corresponding control neuron, c_{ij}, is greater than 0.

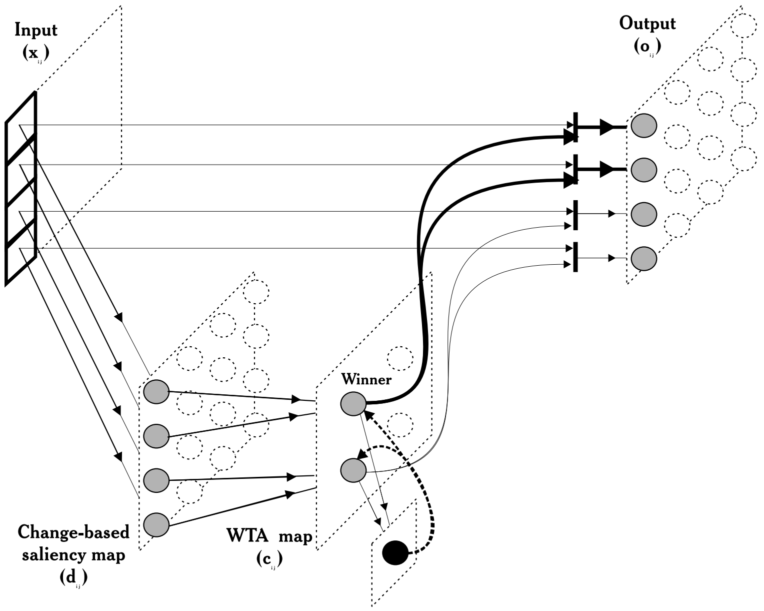

Fig. 2. Attention mechanism. Neurons in the change-based saliency map perform a temporal derivative of the input, and their output drives the activation of the control neurons in the WTA map. These neurons are globally inhibited by an interneuron in charge of selecting a winner, which signals the most prominent region/s of change. Finally, presynaptic facilitation triggered by the WTA map output c_{ij}, allows information about relevant regions to pass to further processing stages.

The segmentation module exploits the general idea that the brain could be using temporal correlations to bind the different features that compound single objects (7; 2). It consists of a network of integrate and fire neurons which are locally connected among themselves via excitatory synapses, and globally connected to a common inhibitory interneuron that sends back inhibition to the whole network. Local excitatory connections provide a fast synchronization of all the neurons responding to a connected region in the input scene, while global inhibition prevents from having disconnected regions oscillating in synchrony. This kind of network is inpired by (11), but here we consider a simplified integrate and fire neuron model (5).

The normalization module is composed by the translation-invariant network, and the scale/orientation-invariant network. The image given by the segmentation module is fed into the translation-invariant module, which is composed of a 2D network of excitatory neurons and four control neurons which modulate the intrinsic recurrent connectivity of the network (see figure 3). The “up” control neuron activates the synapses between each pixel neuron and its upper

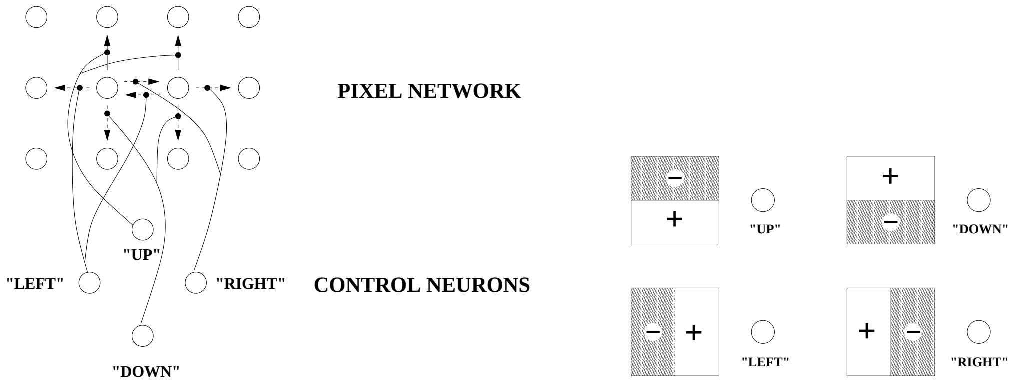

Fig. 3. Traslation-invariant network. Left: each neuron represents a pixel of the image given by the segmentation module, and excitates its 4 nearest neighbours. Each one of the four control neurons activates a particular connectivity (up, down, left or right). Right: receptive fields of the control neurons.

neighbour. The “down” control neuron activates the synapses between each pixel neuron and its bottom neighbour, and analogously for the “left” and “right” control neurons. These neurons compete among themselves through a WTA mechanism. Therefore, when the “up” control neuron wins, the network performs a translation of 1 pixel in the “up” direction. The receptive fields of these control neurons are shown in figure 3. For example, the “up” neuron receives excitation from the lower pixels, and inhibition from the others. Therefore, this neuron is activated as long as there are more pixels in the lower part of the image than in the upper one. The system is then moving the image until its “center of gravity” coincides with the origin. Then, no control neuron wins the competition and the computation is finished.

The output of this network is fed into the scale/orientation-invariant network. The connectivity between these two networks can be easily chosen so that the information is now represented in polar coordinates (see figure 4). The structure and dynamics of this network is exactly the same as in the translation-invariant network, including the control neurons. Since the dilation/contraction operations, as well as the rotation operations, can be described as translations in polar coordinates, this network is performing normalization with respect to both size and orientation.

Finally, the output from the normalization module is fed into the recognition network. This module is composed by two layers of neurons. The first one develops internal representations in order to recognize previously shown stimuli. The second module learns associations between these representations and the label of the stimulus (”circle”, ”square” or ”triangle”), which is given by the teacher. The dynamics of this module is biologically realistic, for details see (5).

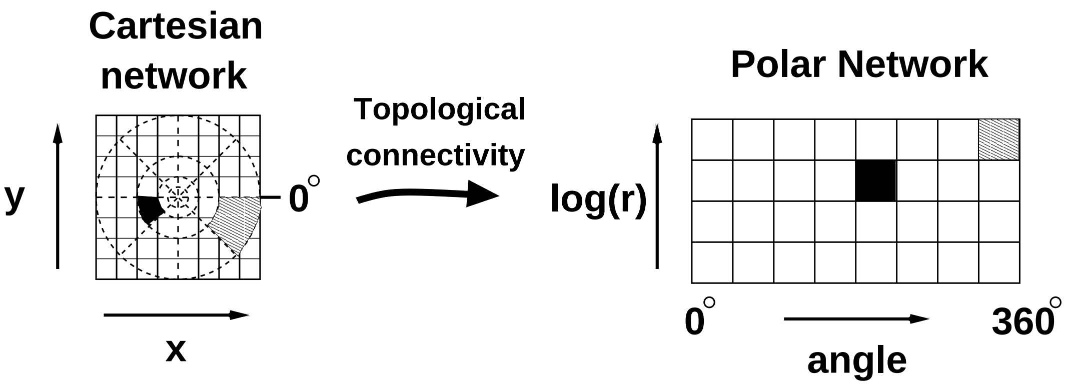

Fig. 4. Scale-invariant network input. The information in the translation-invariant network is represented in cartesian coordinates. The connectivity to the scale-invariant network is adjusted so that the information is now represented in polar coordinates. This connectivity is local and topological.

3 Results and discussion

To test the model, we stimulated the network with sequences of input images consisting of black objects on a white background. We used different size versions of each object, which were presented at various locations, with simultaneous appearance of up to 4 objects allowed. The succesive processing steps achieved attention fixation on changing regions; segmentation of the different objects that appear in these regions; normalization to standard size, position and orientation; and finally recognition. An example of the output from the different modules for a typical trial is shown schematically in figure 5.

The attention module was inspired by the pioneering model of Koch and Ullman (4) and its extensions (3). Our approach could be integrated in these models in order to enhance their functionality by allowing transients processing. The change-based saliency map, and its associated WTA map, could be naturally understood as a new feature map to be added to the color, orientation and intensity maps. These maps could be integrated in a unique saliency map in charge of encoding not only static patterns, but also animated image sequences.

The routing complexity problem can be stated as the incredible amount of control neurons needed in classical routing networks (8). Contrary to these feedforward models, our approach introduces recurrent connections, which keep the system very simple, as well as closer to biology. This solution drastically reduces the number of required control neurons down to 8.

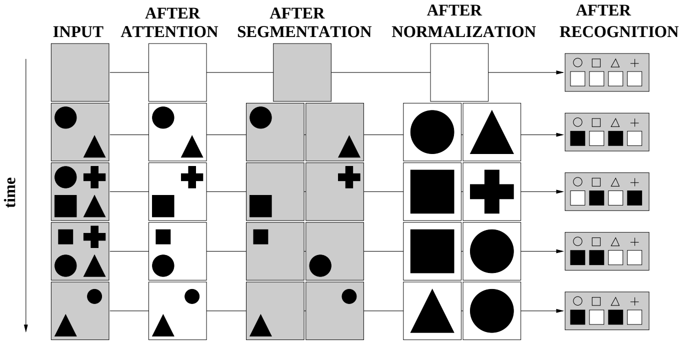

Fig. 5. Schematics of the network processing on a typical input sequence.

References

[1] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (1980) 193-202.

[2] C. M. Gray, Synchronous oscillations in neuronal systems: mechanisms and function. J. Comput. Neurosci. 1 (1994) 11-38.

[3] L. Itti, C. Koch, Computational modelling of visual attention, Nature Reviews. 2 (2001) 1-9.

[4] C. Koch, S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, Hum. Neurobiol. 4 (1985) 210-227.

[5] L.F. Lago-Fernández, M.A. Sánchez-Montañés, F.J. Corbacho, A Biologically Inspired Visual System for an Autonomous Robot. Neurocomputing 38-40 (2001) 1385-1391.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural comp. 1(4) (1989) 541-551.

[7] C. von der Malsburg, The correlation theory of brain function, in: E. Domany, J.L. van Hemmen and K. Schulten, eds., Models of Neural networks II (Springer, Berlin, 1994) 95-119.

[8] B.A. Olshausen, C.H. Anderson, D.C. Van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. 13(11) (1993) 4700-19.

[9] M. B. Reid, L. Spirkovska, E. Ochoa, Simultaneous position, scale, and rotation invariant pattern classification using third-order neural networks, International Journal of Neural Networks - Research and Applications. 1 (3) 154-159.

[10] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neurosci. 2(11) (1999) 1019-1025.

[11] D. Wang, D. Terman, Locally excitatory globally inhibitory oscillator networks. IEEE Transactions on Neural Networks 6(1) (1995) 283-286.

[12] L. Wiskott, How does our visual system achieve shift and size invariance?, in L. Van Hemmen and T. J. Sejnowski (Eds.) Problems in Systems Neuroscience, Oxford University Press, 2001.