Using Multiple Self-Organizing Architectures for Object Recognition

Alessio Plebe and Rosaria Grazia Domenella
Department Cognitive Science – University of Messina
Messina, ITALY
{aplebe,rdomenella}@unime.it

Abstract - *This work proposes a model of the development of visual object recognition, based on the combination of two different artificial neural architectures, both supporting self-organization: LISSOM and SOM. The former is a better approximation of the biological computations in cortical areas, including lateral connections, the latter is best suited for a simple synthesis of non localized processes, like object categorization.*

Key words - object recognition, self-organizing maps, visual invariance

1 Introduction

Despite being vision largely the most studied function of the brain, with a huge collection of neuroscientific and neurocomputational studies on early vision, object recognition remains scarcely investigated and poorly understood yet. This gap is certainly not marginal since for primates object recognition is the most valuable outcome of the visual system. Neurocomputation can offer a methodology in shedding light inside natural vision by simulating empirically assessable brain computations [17]. In this context a fertile concept has been self-organization, applied in the first mathematical model able to simulate the spontaneous development of important features in brain visual areas [23]. Since then self-organization has been the basis of further development in modeling vision [18]. A promising architecture is the LISSOM (*Laterally Interconnected Synergetically Self-Organizing Map*), where in a simple formulation the main neural mechanisms of a cortical map are included: Hebbian plasticity, competitive constrains, intercortical excitatory and inhibitory connections [19, 1].
This progress has been fruitful for several areas of vision, less for object recognition, a reason is the scarce empirical knowledge of the relevant cortical functions. In a well-known review Farah and Aguirre [5], comparing PET and fMRI studies on the neural substrates of human visual recognition, concluded that "The pooled results of these studies can be summarized by the following, rather anticlimactic, statement: visual recognition activates posterior brain regions.".
The direction proposed in this work is to combine neurocomputational architectures close to realistic cortical computations with the more abstract SOM architecture suitable to synthesize brain functions difficult to localize in specific areas. Moreover, using the SOM as a conceptual final map in the model allows an easy combination of non visual inputs.
We believe that a main reason of the difficulty in understanding human visual recognition, and

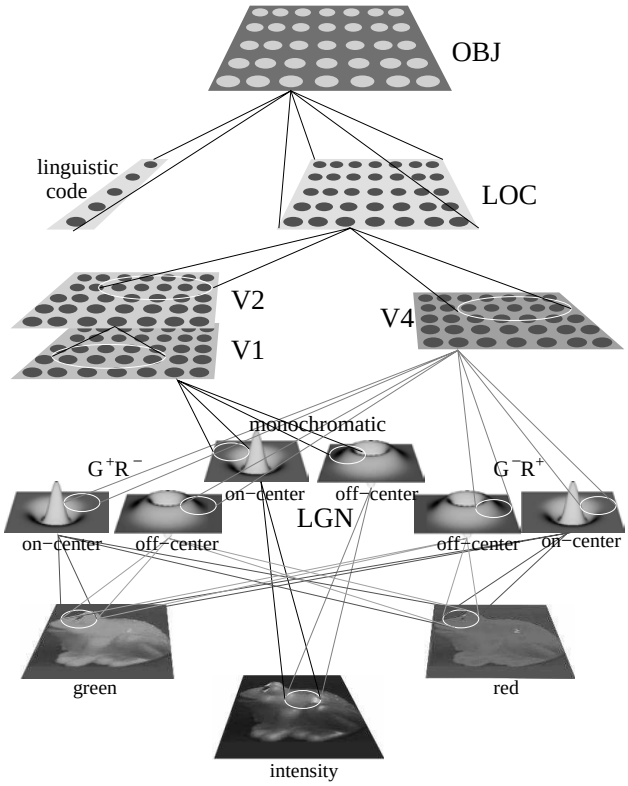

Figure 1: Scheme of the model architecture. The processing pathway is from the bottom to the top of the figure.

the cause of its spread progressing in the brain, is the intimate relationship with the linguistic organization of meaning [13]. Therefore, the model here proposed includes an encoding of linguistic stimuli coincidental with the pure visual signals, representing the typical ostensive naming condition of newborns. The two architectures are clearly faithful in different degree in simulating actual brain computations, however they share the same focus on reproducing the emergence of functions from input patterns, and therefore are both good candidates for an exploration on how object recognition emerges in humans, which is the main purpose of this model.

2 Modeling the development of visual recognition

Most computational models of recognition are aimed at reproducing the performances of the mature visual system [6], this is common also in models based on neural networks [14]. We believe that to understand how the brain areas involved in recognition gradually succeed in developing their final functions would be a major key in revealing how humans can recognize objects. There is a large evidence that the visual system develops thanks to epigenetic interactions, even before eye disclosing [9, 10]. Moreover, it is known from developmental psychology that perception of object entities take place in a period between 8 and 16 months [12], where object naming constitute an important external input [15]. At 24 months even a single ostensive naming can induce correct visual recognition [3], and at the same age recognizing named object seems to involve more visual mechanisms, like shape processing, compared to unnamed objects [20].

The model here proposed is made of several artificial cortical layers, where all layers closer to the eye's input are of LISSOM type. The last map is a SOM, and does not correspond to any specific area in the brain, it is a more abstract module synthesizing a function that certainly involves many loci of the brain, probably extending also beyond the occipital and temporal lobes, like in the prefrontal and in the perirhinal cortex. The overall scheme is depicted in Fig. 1. There is a visual input, split into two color planes and an intensity plane, and the linguistic input, coded into the SOM input vector.

2.1 The architecture for simulating cortical areas

In the LISSOM map each neuron is not just connected with the afferent input vector, but receives excitatory and inhibitory inputs from several neighbor neurons on the same map. The activation level a_i of a neuron i at a certain time step k is be given by:

$$a_i^{(k)} = f\left(\gamma_{\mathrm{X}}\vec{x}_i \cdot \vec{v} \; + \gamma_{\mathrm{E}}\vec{e}_i \cdot \vec{y}_i^{(k-1)} \; + \gamma_{\mathrm{H}}\vec{h}_i \cdot \vec{z}_i^{(k-1)}\right), \tag{1}$$

where the vectors $\vec{y}_i$ and $\vec{z}_i$ are the activations of all neurons in the map where exists a lateral connection with neuron i of, respectively, excitatory or inhibitory type. Vectors $\vec{e}_i$ and $\vec{h}_i$ are composed by all connection strengths of, respectively, the excitatory or inhibitory neurons projecting to i. The vectors $\vec{v}$ and $\vec{x}_i$ are the afferent inputs and the corresponding synaptic efficiencies. The scalars γ_{X}, γ_{E}, and γ_{H}, are constants modulating the contributions. The map is characterized by the matrices $\mathbf{X}, \mathbf{E}, \mathbf{H}$, which columns are all vectors $\vec{x}$, $\vec{e}$, $\vec{h}$ for every neuron in the map. The function f is any monotonic non-linear function limited between 0 and 1. The final activation value of the neurons is assessed after a certain settling time K. All connection strengths adapt according to the general Hebbian principle, including a normalization mechanism counterbalancing the overall increase of connections. The afferent connections to a neuron i will be modified at each training step by the following rule:

$$\Delta\vec{x}_i = \frac{\vec{x}_i + \eta a_i \vec{v}}{\|\vec{x}_i + \eta a_i \vec{v}\|} - \vec{x}_i, \tag{2}$$

and similarly for weights $\vec{e}$ and $\vec{h}$.

The LISSOM has been adapted as a model for vision [1], with an organization of the components of input as receptive fields. The vector $\vec{v}$ is now made of afferent signals organized in a two dimensional fashion, and $\vec{x}$ can be thought as a two dimensional function shaping the receptive field. Therefore, using two orthogonal indexes r and c, equation (1) may be rewritten as:

$$a_{r,c}^{(k)} = f\left(\gamma_{\mathrm{X}}\vec{x}_{r,c} \cdot \vec{v}_{r,c} \; + \gamma_{\mathrm{E}}\vec{e}_{r,c} \cdot \vec{y}_{r,c}^{(k-1)} \; + \gamma_{\mathrm{H}}\vec{h}_{r,c} \cdot \vec{z}_{r,c}^{(k-1)}\right), \tag{3}$$

where now $\vec{v}_{r,c}$ is a vector composed by all values in a two-dimensional array, included in the circular receptive field projected by the neural element $x_{r,c}$. There is a topological correspondence between a translation of r, c on the map and the translation of the field in the input array.

As seen in Fig. 1 LISSOM maps are named in analogy of the corresponding cortical areas. There are also lower maps called LGN, with relation to the biological Lateral Geniculate Nucleus, with receptive field shaped by a classical "Mexican-hat" function, acting as "on-center" and "off-center" cells, and color opponent cells [21].

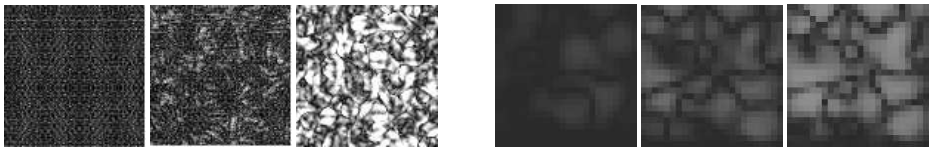

Figure 2: Development of organizational domains in V1 (left) and V4 (right). The selectivity to orientation (left) or hue constancy (right) is shown in gray scale.

2.2 The architecture for the abstract final map

The final map of the model is the well known SOM of Kohonen [11], where the learning rule is on a *winner-take-all* basis. The map is made by M neurons, arranged in a two-dimensional coordinate r, each associated with a vector $\vec{x} \in \mathbb{R}^N$ where N is the dimension of the input data vectors $\vec{v}$. Presenting an input v to the map there will be a winner neuron w satisfying:

$$w = \arg \min_{i \in \{1,\ldots,M\}} \{\|\vec{v} - \vec{x}_i\|\}. \tag{4}$$

Once identified the winner, during training the neural vectors are updated using:

$$\Delta \vec{x}_i = \eta e^{-\frac{\|\vec{r}_w - \vec{r}_i\|^2}{2\sigma^2}} (\vec{v} - \vec{x}_i), \tag{5}$$

where η is the learning rate, σ the amplitude of the neighborhood affected by the updating. Both η and σ are decreasing functions of the training epochs. In the computation of (4) the components of $\vec{v}$ encoding the linguistic input are weighted by a parameter increasing with the training epochs, simulating the increase in attention to ostensive naming of objects.

3 Main features of the model

The set of natural images used in the experiment is the COIL-100 benchmark library [16], a collection of 100 ordinary objects, each seen under 72 different perspectives.
The cortical map named V1, in relation with the primary visual area, collects its afferents from the monochromatic sheets pair in the LGN, and is followed by the map V2, which has a lower resolution and larger receptive fields. The main phenomena reproduced by this model in these areas is the development of orientation domains, where many cells are sensitive to a preferred orientation [22]. The training uses artificial elliptical blobs for the first 10000 steps followed by natural images for other 10000 steps, after the initial formation of the orientation domains. This procedure accords with the known role of spontaneous activity in the pre-natal neural development and in the first period after eye opening [7, 2]. In the left side of Fig. 2 the emergence of the orientation domains during the training is shown.
The color path is processed by the LISSOM map V4, named as the biological area especially involved in color processing [24]. The main feature of the cortical color process is color constancy. It has been shown by psychophysical experiments in human infants that this capability is not present from birth either, but develops sometimes between two and four months of age [4]. The right part of Fig. 2 shows the development of domains with constant hue inside the V4 map during the training. At the beginning there is a low sensitivity, peaked in the middle range between red and green. At the end the color sensitivity of all patches is uniformly distributed along the hue range.

type of transformation	image		LOC map	
	avg	stdv	avg	stdv
rotation of 30°	0.781	0.140	0.903	0.102
rotation of 60°	0.648	0.221	0.756	0.181
size downscaling of 80%	0.637	0.119	0.794	0.129
size downscaling of 70%	0.547	0.126	0.655	0.181
translation of 10%	0.463	0.125	0.586	0.158
translation of 20%	0.207	0.133	0.397	0.155

Table 1: Correlations between images with viewpoint transformation (middle column), and the corresponding LOC map (right column). The values are the average over all 100 objects.

The paths from V4 and V2 rejoin in the map LOC, which has larger receptive fields, and correlates with the human LOC (Lateral Occipital Complex) brain area which seems to be strongly involved in object recognition [8]. One of its main feature is a reasonable invariance to size, specific visual cues, and some perspective transformation. The model LOC achieves by unsupervised training, using all COIL-100 images in all possible view, a remarkable invariance with respect to viewpoint, position and size. The quantitative assessment of invariance, visible in Tab. 1 has been obtained by measuring the cross-correlation between images.

The highest map in the model is called OBJ, and is of SOM type. It processes as vector input the whole content of LOC, ignoring the spacial organization of the data, and a vector coding linguistic inputs. The common ostensive naming of object given to infants by adults is simulated by the coding of all objects in a 100-dimension vector. Naming does not pose a normative labeling on objects, it is just a signal coincidental with the visual input. The organization of all objects in the map OBJ is shown in Fig. 3. The figure is obtained by overlapping every neuron in the map with the object for which that neuron is the most frequent winner. In the map coexist several overlapped organizations: by color, by shape, by symmetries, producing a consistent categorization of most objects. The two prevalent ordering criteria of OBJ, shape and color, are shown separately in Tab. 2. In both matrix each neuron of the map is labeled according to the category of shape/color which mostly activate it. If the difference between the prevailing category and the second one is not significant the neuron is left blank. From experiment without the linguistic component in the OBJ input

3	3	6	B	A	7	7	5	6		4	4	4	8	8	4
2	C		B	1	C	5	9						8	4	4
5	5	2	9			C	5	9	9	1	8	8	4		6
2			6		C		5	8		8	8		8		2
2	3				7		2	2	8		4	4	7		
5	5	6	6	3	9		2	2	7		B	7			6
5		5			9		2		8	7	7	A	7	7	6
1	1	1	1	5	7	9	2	2	2	7		7		7	6
1	1	1	1	1		3				B	7	7	7		3
5	5	5	1	1	C	C	C	7	7	B			7		3
5		5	5	3	3			C	7	3	3	3	7		9
3		5	5	1	3	2					3	3	3	3	4
3		5	5	5		B	5	6		A		2	3	8	4
4		5	5	5	5	5	3	3		1		2		8	6
9		5		5	5	5	5	2	5		1		6	6	9
2	5	5		2		2	2	2	2		A	5	6	9	9

1	h-parallelepiped
2	round
3	composed
4	q-cylindrical
5	q-h-parallelepiped
6	cylindrical
7	cup-shaped
8	q-v-parallelepiped
9	body
A	conic
B	parallelepiped
C	q-parallelepiped

5	5	A	1	7	5	A	5	3		5	3	3	5	A	A
5	6	A	1	1	6	5	5		5				7		3
5	5	5	5			5	5	5	5	5	5	5	5		4
5	5		4	4	5		5	1		7	5		5		5
5	5				5	5	1	1	1	5	5	5	2		
5	5	2	2	6	1		1	1	5	5	5	2			2
8		5			1		1		1	5	5	5	5	4	1
1	1	1	5	5	1	1	1	1	1	5		5		5	1
A	A	A	1	1	6	6				1	5	5	5		6
1	1	1	1	1	1	6	6	6	A				5		6
4		1	1	1	1			6	2	6	6	6	4		1
A		2	2	5	5	8					6	6	6	6	2
A		2	2	2		5	2	8		1		8	6	2	9
4	4	2	2	2	2	1	1	1		6				2	2
4		4		2	2	1	2	2	2		6	2	2	2	2
4	4	A		6		2	2	2	2		2	2	2	8	8

1	yellow
2	red
3	white-green
4	green
5	white
6	brown
7	white-blue
8	pink
9	white-red
A	blue

Table 2: Layout of objects in the OBJ map, according to their shape (left) and hue (right) characteristics. Shape prefixes: “h”=horizontal, “q”=quasi, “v”=vertical.

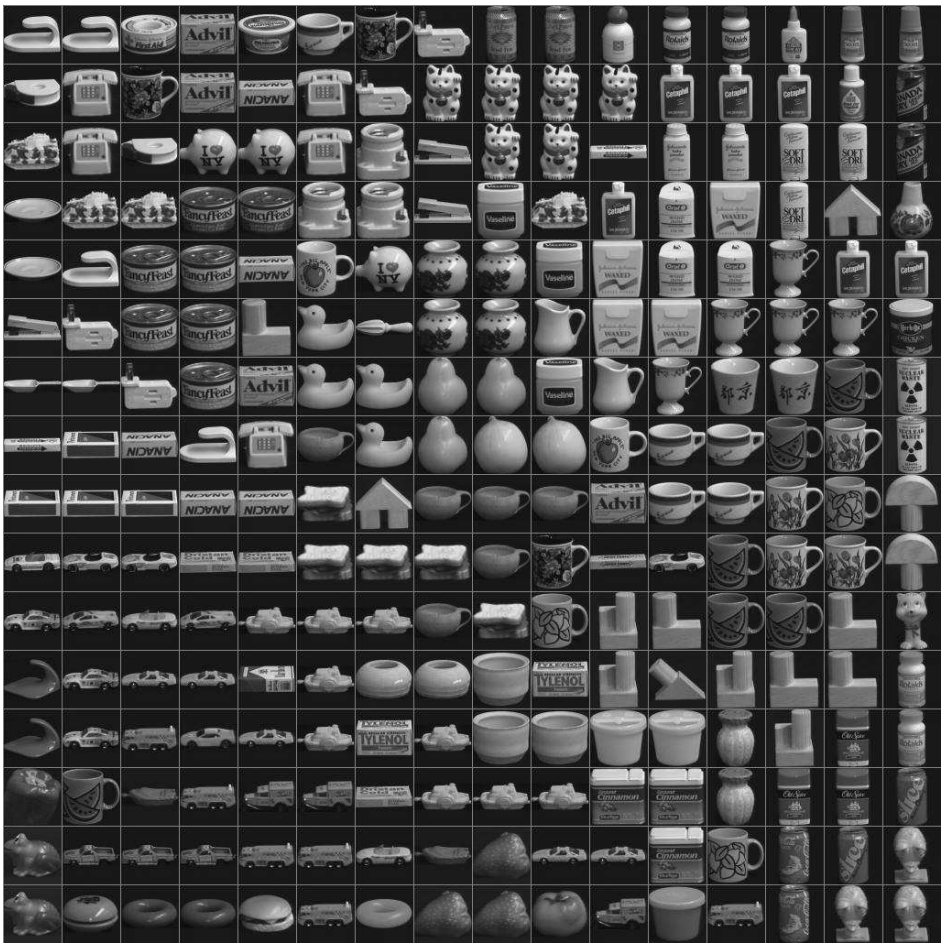

Figure 3: Organization of objects in the OBJ map of the model. Each neuron of the map is labeled using the base view of the object prevailing on that neuron.

it has been observed that the most significant influence of the linguistic component is the concentration of activities induced by the same object in different views. Without language, the average spread is on 9.6 neurons with an area of 48 units, in combination with language it is reduced to 7.8 neurons in an area of 34 units. Figure 4 displays the difference between activations with and without the language contribution.

4 Conclusions

A model for simulating the emergence of visual recognition has been introduced. It attempts to overcome the gap of neuroscientific knowledge between lower and upper visual functions with the combination of the two self-organizing artificial architectures LISSOM and SOM, levering on their different features: a better biological plausibility for LISSOM and a wider generality for the SOM. Nevertheless, being the model aimed at simulating a very complex cognitive function, there are many limitations with respect to the human object recognition. A strong simplification is in segregating processes in single modules, like color in V4, and in neglecting backprojections in a pure hierarchical model. Also the treatment of the linguistic contribution is drastically simplified. Despite these limitations, the model can reproduce

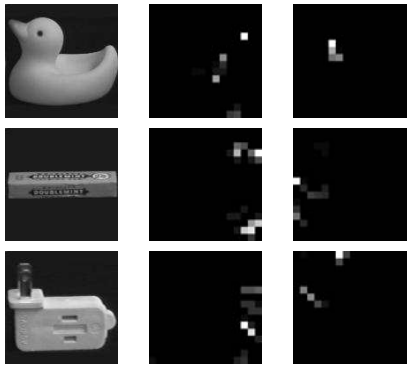

Figure 4: Spread of the 72 views of an object in the OBJ map without linguistic input (middle) and with linguistic input (right) for some objects (left). The brightness of units in the map is proportional to the number of hits.

several mechanisms essential for visual recognition, without any explicit modeling of the processing functions necessary for this goal, and looks like an interesting direction for further investigations, with the inclusion of some of the missing pieces just mentioned.

References

[1] J. A. Bednar. *Learning to See: Genetic and Environmental Influences on Visual Development*. PhD thesis, University of Texas at Austin, 2002. Tech Report AI-TR-02-294.

[2] B. Chapman, M. P. Stryker, and T. Bonhoeffer. Development of orientation preference maps in ferret primary visual cortex. *Journal of Neuroscience*, 16:6443–6453, 1996.

[3] E. V. Clark. What's in a word: On the child's acquisition of semantics in his first language. In T. E. Moore, editor, *Cognitive development and the acquisition of language*. Academic Press, New York, 1973.

[4] J. L. Dannemiller. A test of color constancy in 9- and 20-weeks-old human infants following simulated illuminant changes. *Developmental Psychology*, 25:171–184, 1989.

[5] M. J. Farah and G. K. Aguirre. Imaging visual recognition: PET and fMRI studies of the functional anatomy of human visual recognition. *Trends in Cognitive Sciences*, 3:179–186, 1999.

[6] I. Gauthier, P. Williams, M. Tarr, and J. Tanaka. Training greeble experts: A framework for studying expert object recognition processes. *Vision Research*, 38:2401–2428, 1998.

[7] I. Gödecke and T. Bonhoeffer. Development of identical orientation maps for two eyes without common visual experience. *Nature*, 379:251–254, 1996.

[8] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41:1409–1422, 2001.

[9] L. C. Katz and E. M. Callaway. Development of local circuits in mammalian visual cortex. *Science*, 255:209–212, 1992.

[10] A. Kirkwood and M. F. Bear. Hebbian synapses in visual cortex. *Journal of Neuroscience*, 14:1634–1645, 1994.

[11] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[12] L. I. Leushina and A. A. Nevskaya. Development of vision and visual notions in infants. *Journal of Evolutionary Biochemistry and Physiology*, 39:67–76, 2003. English translation from *Zhurnal Evolyutsionnoi Biokhimii i Fiziologii*.

[13] D. Marconi. *Lexical Competence*. MIT Press, Cambridge (MA), 1997. Ediz. it. *Competenza Lessicale*, Laterza, 1999.

[14] B. W. Mel. SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[15] D. L. Mills, S. A. Coffey-Corina, and H. J. Neville. Language comprehension and cerebral specialization from 13 months to 20 months. *Developmental Neuropsychology*, 13:397–445, 1997.

[16] H. Murase and S. Nayar. Visual learning and recognition of 3-d object by appearence. *International Journal of Computer Vision*, 14:5–24, 1995.

[17] E. Rolls and G. Deco. *Computational Neuroscience of Vision*. Oxford University Press, Oxford (UK), 2002.

[18] L. Schwabe and K. Obermayer. Modeling the adaptive visual system: a survey of principled approaches. *Neural Networks*, 16:1353–1371, 2003.

[19] J. Sirosh and R. Miikkulainen. Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, 9:577–594, 1997.

[20] L. B. Smith. Children's noun learning: How general learning processes make specialized learning mechanisms. In B. MacWhinney, editor, *The Emergence of Language*. Lawrence Erlbaum Associates, Mahwah (NJ), 1999. Second Edition.

[21] R. L. D. Valois and G. H. Jacobs. Primate color vision. *Science*, 162:533–540, 1968.

[22] W. Vanduffel, R. B. H. Tootell, A. A. Schoups, and G. A. Orban. The organization of orientation selectivity throughout the macaque visual cortex. *Cerebral Cortex*, 12:647–662, 2002.

[23] C. von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kibernetic*, 14:85–100, 1973.

[24] S. Zeki. Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelenghts and colours. *Neuroscience*, 9:741–765, 1983.