Towards an Inclusive Computational Model of Visual Cortex

M. Gheiratmand, *Student Member, IEEE*, H. Soltanian-Zadeh, *Senior Member, IEEE,* and H. Khaloozadeh

***Abstract*—Understanding the primates' visual system has been one of the challenging problems of different groups of scientists for years. Though many studies, from physiology and neuroscience to computer vision, are done on different aspects of visual processing in the cortex, a comprehensive computational model of visual cortex is still missing. We have implemented a computational model of object recognition in ventral visual pathway in our previous work. This hierarchical model covers visual areas V1/V2, V4/PIT, and AIT sending inputs to the Prefrontal Cortex (PFC) for categorization. To extend our model, in this work, we have added a simple model of motion detection in neurons of areas V1 and MT of the dorsal stream to our previous model. This has enabled the model to perform another principal function of the visual cortex, i.e., motion perception.**

I. Introduction

IT is confirmed that cells in different areas of the visual cortex are specialized for different types of visual information, such as motion, form, and color, and have different properties. Ungerleider and Mishkin (1982) showed that the visual information is processed in two separate pathways. The ventral pathway extends from V1 to inferior temporal cortex (IT), including V4, and is known to perform object recognition. The dorsal pathway begins in V1 and turns upward to the posterior parietal area, including the middle temporal area (MT) and is responsible for location and motion perception [1].

However, some scientists believe that computational architecture of the cerebral cortex is very similar from one neocortical area to another, but the inputs to every cortical area is quite different. This idea is successfully used in motion analysis model of [2]. The object recognition model of [3] is also based on an extension of the organization of simple and complex cells in the striate cortex of the primates, proposed by Hubel and Wiesel (1968).

Towards construction of a complete mathematical model of visual processing in the visual cortex, we combine biologically motivated models of different visual cortical areas. First, the extended HMAX object recognition model [3] is implemented to simulate the function of neurons in areas V1/V2, V4/PIT, and AIT along the ventral pathway. Some enhancements and simplifications are done on the model to improve the processing time and performance of the model. In the next step, we add a simple model of motion detection in the areas V1 and MT to extend our model of visual cortex. In this work, only component direction selective cells are modeled using the energy model of [4]. For this part, we use parameters which are consistent with those defined for the object recognition model.

Other object recognition systems are investigated in [3] and [5], but they all lack the simultaneous physiological plausibility and high performance, achieved by the extended HMAX model proposed in [3], [5], [6]. The fact that parameters in this model are not tuned to obtain the optimal performance, but are tuned to match the physiological properties of the neurons in the corresponding areas [5], [6], along with the high performance of model to input images of the real world, make it a solid frame to construct our inclusive model on it. Some works are done to add feedbacks and attention mechanisms to the HMAX model [7]. Though achieving good performance and decreasing the response latency, [7] uses only the V1 layer of the HMAX model, which makes it farther from the biological visual system.

Considerable works are done on motion analysis in the cortex by Heeger, Movshon, Simoncelli and others. Movshon proposed a two stage hypothesis for the motion analysis by neurons in area MT [1], [8]. Some neurons in area V1, like most of the neurons in MT are component direction-selective, responding only to components of a pattern moving preferably perpendicular to their orientation axis [1], [9]. About 20% of the MT neurons are pattern direction-selective which compute the global motion of the object [1]. These cells receive inputs from the component cells of areas V1 and MT. In [8], both direction-selective cells in V1 and direction- and speed- (velocity) selective pattern cells of MT are modeled using a similar sequence of operators including: linear filtering, half-square rectification, and divisive normalization. The cascade model described in [9] is a relative model of [8] fitted to the responses of individual MT neurons.

In this work, we do not address the binding problem, as the visual information is processed through separate paths, which receive separate types of inputs. In future work, we may use a common input for both processes.

Manuscript received July 5, 2008.

M. Gheiratmand is with the Electrical Engineering Department, K.N. Toosi University of Technology, Tehran 16315-1355, Iran (phone: 98-21-22674284; fax: 98-21-88785081; e-mail: m.gheiratmand@ ee.kntu.ac.ir).

H. Soltanian-Zadeh is with the Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, University of Tehran, Tehran 14395-515, Iran, School of Cognitive Sciences, Institute for Studies in Theoretical Physics and Mathematics (IPM), Tehran, Iran, and Image Analysis Laboratory, Radiology Department, Henry Ford Health System, Detroit MI, 48202 USA (e-mails: hszadeh@ut.ac.ir, hamids@rad.hfh.edu).

H. Khaloozadeh is with the Electrical Engineering Department, K.N.Toosi University of Technology, 16315-1355 Tehran, Iran (e-mail: h_khaloozadeh@kntu.ac.ir).

II. Model Implementation

Fig. 1 shows a block diagram of the proposed model along with the corresponding areas of each block in the visual cortex.

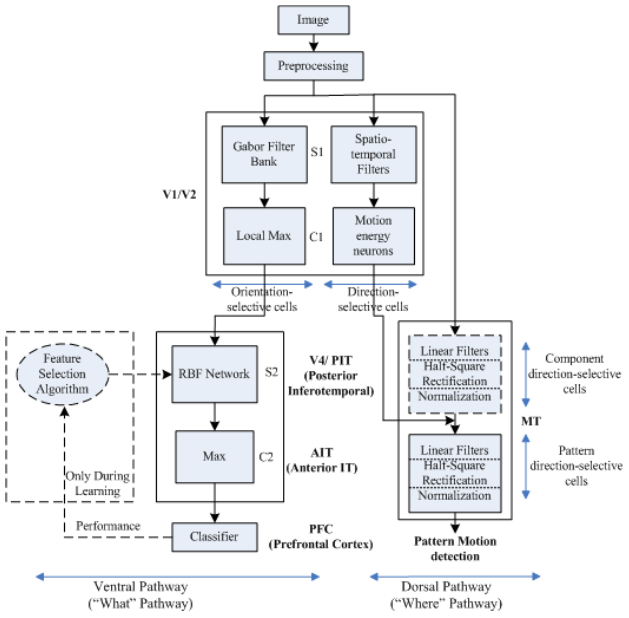

Fig. 1. Block diagram of the visual processing model. The corresponding visual areas of each part are included.

A. Object recognition model of ventral pathway

The object recognition model proposed in [3] is a feed-forward hierarchical quantitative model of object recognition that accounts for the very first milliseconds of processing in the ventral stream of primate visual cortex. In its simplest form, the model has four layers: S1, C1, S2, and C2. The model extracts shift- and scale-invariant features from an input image and sends it to a trained classifier to decide about its object category. In this work, we have trained the model for two objects, but the model is tested on 101 object categories [3], [5]. Details of the implemented HMAX model are described in Appendix.

Our investigations on the role of each C1 frequency band, shows that the existence of all the bands necessarily do not improve the classification performance. We have also investigated the effect of using different patch sizes to produce S2 RBF neuron centers. In general, features of intermediate sizes work better; because, compared to larger patches, they are more flexible in matching a greater number of inputs, and compared with smaller patches, they are more selective to the desired object [10]. To run the feature selection algorithm, we have used the first scale band of C1, corresponding to the first two scale bands of S1 cells. We have also executed the algorithm on the S2 features produced by using the second patch size (8×8).

1) Feature selection: Features in Layer S2 are chosen randomly from a set of positive images. As a result some features might not be useful for classification, and increases the response latency at the same time. To lessen these drawbacks we have applied a sequential backward feature selection algorithm to the randomly chosen prototypes of Layer S2. We have omitted 30% of the worst features from the primary set of 100 S2 neurons. This, in addition to reducing the processing time, has led to an increase in the classification performance. Other methods such as k-Means clustering were also examined to divide the S2 features in two useful and bad features. Due to the highly random characteristic of the features, no reasonable results were obtained. As the output values of the model are either 1 or 0 (belonging or not belonging to an object category) training the RBF network is also not a practical method of improving the S2 prototypes.

B. Motion detection model of dorsal pathway

We have used the motion energy model of Adelson and Bergen [4], to provide our model with the ability of motion analysis. This model simulates the characteristic of some cells in area V1, and most of the cells in area MT of the dorsal pathway. These cells are component direction-selective, showing response to motion in a specific direction [1]. Pattern direction-selective cells in the area MT, which are selective to the general motion of the object, receive input from the component direction-selective cells of V1 and MT.

By now we have modeled V1 orientation-selective cells, having spatial receptive fields. We should now produce spatiotemporal receptive fields. These receptive fields are simply produced by multiplication of temporal and spatial impulse responses. These kinds of separable spatiotemporal responses are physiologically and psychophysically plausible [4]. The motion energy model implementation is described in the following stages:

1) Spatial filters: We have used pairs of sine and cosine Gabor functions with the same properties used in the modeling of ventral stream as the spatial response of the cells. This is done in order to stay consistent with the object recognition model and its physiologically tuned parameters. Reference [4] has used the second and third Derivatives of Gaussian to produce a pair of quadrature spatial filters. Though a sine and cosine pair of Gabor functions are often considered as a quadrature pair, we have removed the DC

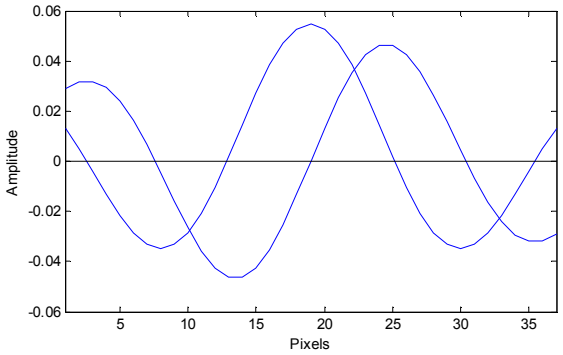

Fig. 2. A quadrature pair of spatial Gabor filters with biologically tuned parameters.

component of the cosine Gabor function in order to produce a refined quadrature pair. Fig. 2 shows a sample quadarture pair of Gabor filters. Details of the Gabor function is described in Appendix.

2) Temporal filters: we have used linear temporal filters introduced in [4], which are said to be plausible approximations to filters inferred psychophysically. These filters have the form of (1) where *n* takes values of 3 and 5. Temporal filters are actually weighting functions that combine the spatial responses of cells in the past to produce the response at the present moment.

$$f(t) = (kt)^n \exp(-kt)[1/n! - (kt)^2/(n+2)!]. \quad (1)$$

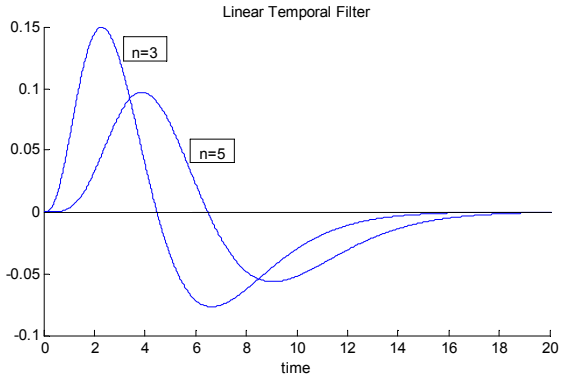

Fig. 3. Two temporal filters duplicating the temporal responses of direction selective cells in areas V1 and MT.

3) Spatiotemporal Filters: These kind of spatiotemporal cell responses are produced by multiplying spatial and temporal filters with the above mentioned forms.

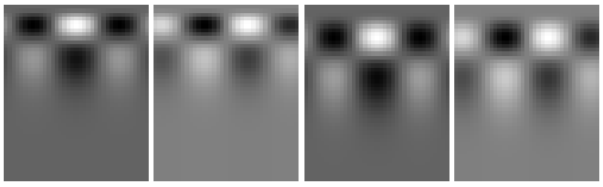

Fig. 4. Spatiotemporal filters produced by multiplication of the sample spatial and temporal filters.

4) Direction selective Spatiotemporal filters: The filters illustrated in Fig. 5 are produced by summation and subtraction of the spatiotemporal filters of Fig. 4.

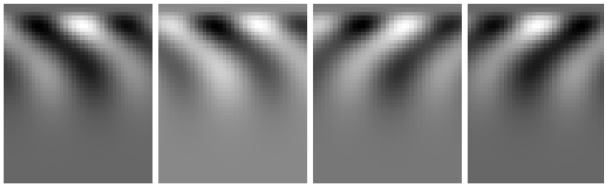

Fig. 5. A pair of leftward and rightward direction-selective spatiotemporal filters.

Each quadrature pair of leftward and rightward filters operates on the spatiotemporal input. The outputs of filters in each pair are then squared and added to obtain the oriented energy in right and left direction as in Fig. 6 (c) and (d). To construct a unit that shows motion in both directions (an opponent energy unit), the outputs of the oriented energy modules are subtracted. In the output of an opponent energy unit, the light parts indicate the motion in rightwards and the dark parts show leftward motion as in Fig.6 (b).

5) pattern motion selective cells: As the pattern direction-selective model of Simoncelli-Heeger [8] is proved to be one of the best models which correctly matches the responses of a large fraction of cells in area MT [11], we are about to use a model based on [8] to make our system pattern motion sensitive. This architecture is biologically plausible and can compute velocity (both speed and direction). The final block in the MT box represents this part.

By modeling direction-selective cells, and adding them to the V1 part of the model, we now have a more complete representation of the cells in area V1, which is proved to carry out both modes of computation i.e. MAX and energy model [12].

III. Experiments and Results

A. Stimulus

To test the object recognition model, we have used *Motorbikes* and *Background* datasets from the CalTech5 image database available at:

http://www.robots.ox.ac.uk/~vgg/data/data-cats.html.

Ten sets of randomly selected train and test images, including 90 (40 positive and 50 negative) and 100 (50 positive and 50 negative) images respectively, were used and the classification performance was averaged over these sets. All images are inverted to grayscale and resized to 140 pixels in height as in [3].

The inputs to the motion sensitive model were spatiotemporal representations of moving bars or edges, such as those shown in Fig. 6 (a).

B. Results

The classification performance of the object recognition model with 16 S1 scales, 4 sizes of patches, and 100 S2 features has been 96%. For two S1 scales, one size of patch (8×8), and selected S2 features, we have achieved the performance of 95% along with a considerable decrease in the processing time. Using similar parameters without S2 feature selection, results in the classification performance of 92%.

Additional experiments were also done to investigate the sensitivity of the extracted C2 features of the object recognition model. Gaussian noise with variances of 0.01 and 0.1 corresponding to SNR=5 and SNR=0.5 were added to the input images. (SNR is calculated as the ratio of the signal power to noise power: $SNR = P_{Sig}/P_N$.) The results showed that, addition of noise with variances of 0.01 and 0.1 results in a considerable decrease of 10% and 14% in the classification performance. The performance of the SVM classifier on the features extracted directly from the input images using PCA was quite robust; no change for SNR=5,

and a 1.15% decrease for SNR=0.5 was observed.

The output of the motion energy model in area V1 and the component cells of area MT are illustrated in Fig. 6. The output of the opponent energy mode in Fig. 6 (b) shows the motion in both right-ward and left-ward directions, while the oriented energy responses in Fig. 6 (c)–(d) can only detect motion in either right-ward or left-ward direction.

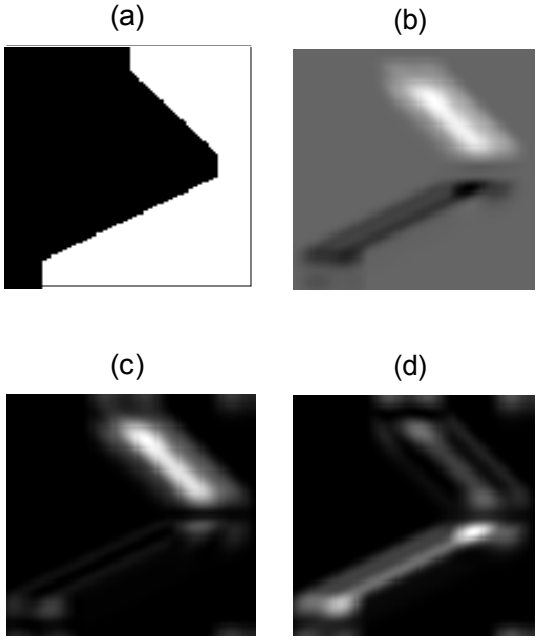

Fig. 6. (a) A stimuli for the motion detection model, consisting of a moving edge presented in the x-t spatiotemporal space. (b) The motion opponent energy output. The light parts demonstrate motion in rightward and the dark parts demonstrate the leftward motion. When the edge is stationary, the response equals zero. (c) The output of rightward motion energy unit. This unit does not cover the leftward motion. (d) The output of the leftward motion energy unit.

IV. Conclusion

In this work, we have combined a model of object recognition in the ventral visual pathway with a simple model of component motion detection in the dorsal pathway and presented a block diagram of the processes executed in the cells of each biological area.

This work is to be extended by adding a pattern motion sensitive model of area MT to the block diagram presented. Also, the inputs to the dorsal and ventral pathways of the model, which were different in this work, are to be united in the form of a moving complex pattern. In addition, the computations and the hierarchical structure used in this work can build a framework for modeling other (visual) cortical areas and provide the model with other abilities such as color perception, which mainly occurs along the ventral pathway. Feedbacks from higher areas to the lower areas in each visual stream, and the interconnecting signals between areas of these two are also of considerable importance in constructing a more inclusive model of the visual cortex.

To integrate quantitative models of different areas of the visual cortex, the parameters of different parts should be adjusted in a way that they preserve consistency with each other. To this end, the parameters in each part of the model can be tuned based on the psychophysical and physiological properties of the neurons in that visual area.

Appendix

The details of the object recognition model and some additional notes about extending the V1 units for the motion sensitive model are described below.

S1 Layer

Spatial receptive fields of simple cells in the primary visual cortex are represented with a bank of Gabor filters as [3], [6]:

$$G_j^i(x.y) = \exp\left(-\frac{\left(x_r{}^2 + \gamma^2 y_r{}^2\right)}{2\sigma_i{}^2}\right) \times \cos\left(\frac{2\pi}{\lambda_i} x_r + \varphi\right). \quad \forall x, y < s_i \tag{2}$$

$$i = 1,...,16. \quad s_i = 7,9,...,35,37.$$

$$j = 1,...,4. \quad \theta_j = 0,45^o,90^o,135^o.$$

$$\begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad {}^o. \tag{3}$$

where θ_j is the preferred orientation of the cells, and s_i is size of the filters ranging from 7×7 to 37×37 which is equivalent to a visual angle of $0.19^o - 1.06^o$ [13]. Aspect ratio γ is equal to 0.3, and the relation between the effective width σ and the wavelength λ with the filter size is empirically optimized as [6]:

$$\begin{cases} \sigma_i = 0.0036 \times s_i{}^2 + 0.35 \times s_i + 0.18 \\ \lambda_i = \dfrac{\sigma_i}{0.8} \end{cases} \tag{4}$$

where φ is the phase offset which is set to zero for the spatial orientation-selective receptive fields. For direction-selective V1 cells, φ takes two values of 0 and 90 degrees to produce a quadrature pair for the motion energy model.

C1 Layer

C1 units correspond to V1 complex cells and combine the outputs of the S1 units with a MAX operation. For each orientation, the maximum operator acts on the output of every spatial frequency scale of the S1 units with a grid cell of size 8×8 to 22×22 with the steps of 2. This operation results in producing invariance to the position of the object (or object parts such as edges) in the image. C1 units also

take maximum over every two adjacent frequency scales in order to make the resulting features scale- invariant.

S2 Layer

This layer is a set of RBF neurons that compute the difference of the calculated C1 features from a new input, with the fixed prototypes set as the centers of the neurons. These prototypes are once set in the learning stage by sampling from a set of 100 positive images. Sampling is done using patches in four sizes of 4×4 to 16×16 with steps of 4, which are placed on random positions in the C1 images extracted from each input image.

C2 Layer

This layer takes maximum over the outputs of the S2 units. As a result a shift- and scale-invariant, and object-selective feature vector with dimension equal to the number of S2 neurons is produced for each input image.

Classifier

The extracted C2 features from train and test image sets are passed to a classifier to be trained. Like the function of area PFC, the classifier can then decide for the category of new input images. We have observed that a linear SVM produced better responses in comparison to kNN (with different numbers of k) and nonlinear SVM. More biologically plausible architectures can be replaced with the SVM classifier [10], which is used here.

REFERENCES

[1] E. R. Kandel, J. H., Schwartz, and T. M. Jessell, *Principles of Neural Science*. New York: McGraw-Hill, 2000, ch. 25, 28.

[2] D. J. Heeger, and E. P. Simoncelli, "Computational models of cortical visual processing," in *Proc. Natl. Acad. Sci*., California, Jan. 1996, vol. 93, pp. 623-627.

[3] T. Serre, L. Wolf, S. Bileschi, M. Reisenuber, and T. Poggio "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411-426, March 2007.

[4] E. H. Adelson, and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am.*, vol. 2, no. 2, pp. 284-299, Feb. 1985.

[5] T. Serre, L. Wolf, and T. Poggio "A new biologically motivated framework for robust object recognition," Massachusetts Inst. of Technology, Cambridge, MA, Tech. Rep. AI Memo 2004-26/CBCL Paper 243, Nov. 2004.

[6] T. Serre, and M. Kouh, "A theory of object recognition: computations and circuits in the feed-forward path of the ventral stream in primate visual cortex," Massachusetts Inst. of Technology, Cambridge, MA, Tech. Rep. AI Memo 2005-036/CBCL Memo 259, Dec. 2005.

[7] E. Bermudez-Contererasl, H. Buxton, and E.Spier "Attention can improve a simple model for object recognition," *J. Image and Vision Computing*, vol. 26, pp. 776-787, 2008.

[8] E.P. Simoncelli, and D.J. Heeger, "A model of neuronal response in visual area MT," *Vision Res.*, vol. 38, no. 5, pp.743-761, 1998.

[9] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon, "How MT cells analyze the motion of visual patterns," *Nature Neuroscience*, vol. 9, no. 11, pp. 1421-1431, Nov. 2006.

[10] T. Serre, J. Louie, M. Reisenhuber, and T. Poggio "On the role of object-specific features for real world object recognition in biological vision," in *Proc. Workshop Biologically Motivated Computer Vision*, 2002, pp. 387-397.

[11] V. Mante, "Testing models of cortical area MT," M.S. thesis, Inst. of Neuroinformatics, ETH/Univ. of Zurich, Switzerland, 2000.

[12] I.M. Finn, and D. Ferster, "Computational diversity in complex cells of cat primary visual cortex**,"** *J. Neuroscience*, vol. 27, no. 36, pp. 9638-9648, Sep. 2007.

[13] T. Serre, "Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines," Ph.D dissertation, Dept. Brain and Cognitive Sciences, Massachusetts Inst. of Technology, Cambridge, MA, 2006.