Abstract—One of the major challenges for the home robots is to recognize objects under different transformations and viewpoints. In this paper, we propose a hierarchical neural network for object recognition inspired by two properties of neurons in the visual cortex: invariance of their responses to stimulus transformations and coding the efficient features in images. To achieve the first goal, we used the trace learning rule to develop neurons with invariance responses to stimulus transformations. For the second goal, we used a model of hierarchical redundancy reduction to extract the most efficient features from images. This hierarchical redundancy reduction uses signals from surrounding neurons in each layer to provide the next layer with most informative features. A set of experiments performed on the coi100 dataset and custom images reveal the high performance of the proposed model for object recognition.

I. INTRODUCTION

HOME applications require robots to deal with different objects from different viewpoints. For example, the robot must locate and recognize objects requested by people such as: bringing a dish from kitchen, locating water jar and closing the door. Such applications require the robot to recognize objects under different viewpoints and backgrounds. Therefore robust object recognition is a principal necessity for home robots. In general, no solution is available to the problem of object recognition. Several methods have been proposed that can recognize objects with high accuracy. However these methods are limited in a variety of conditions: number of objects must be kept small, number of training examples must be large, lighting and view-point should not change, etc. In contrast, the ability of human object recognition is supreme compared to the most successful methods in machine vision. Human can perform object recognition for different view-points, distances, lighting and transformations (affine, 3D).

The structure and function of the primate visual cortex has inspired many models and algorithms in the field of machine vision. Starting with the paper of Hubel and Wiesel in 1962 [1], there has been a large growth of studies of the neuronal selectivity in the visual cortex. These studies have revealed a wealth of information about the processing of images in the cortex. It is now widely believed that object recognition pathway has a hierarchical structure along which the neurons extract features with increasing complexity [2]. Neurons in the primary areas of this hierarchy detect contrast changes in small spatial scales while in higher areas neurons detect complex shapes like hands, faces or objects [1], [3]. Several authors have proposed that neurons in the visual cortex extract features which are statistically optimal for object recognition [4]-[5].

In this paper, we propose a hierarchical model for object recognition in robots based on structure and neuronal properties of the primate visual cortex. Two main properties of the proposed model are the invariance of neuronal responses to object transformations and their selectivity to statistically efficient features. One important property of neurons in the visual cortex is that their responses are invariant to object transformations [6]-[7]. The trace learning rule has been proposed to provide such invariance selectivity for neurons. We use this learning rule in a hierarchical architecture to provide invariance selectivity to 3D object rotations. Another neuronal property in the visual cortex is the coding of locally optimal features for object recognition [4]-[5], [8]. We train the horizontal connections between the neurons of different layers to code the efficient features in each layer. Results of experiments demonstrate the ability of this model to perform robust object recognition under a variety of transformations.

II. LITERATURE REVIEW

Primary models of object recognition used correlation based template matching to detect objects in images [9]. However these models are very sensitive to object transformations and can be used only in controlled environments. In order to overcome this shortcoming, component based models have been proposed that extract features from objects and use them for object recognition [10]-[12]. For these models, a trade-off between selectivity and invariance is unavoidable; the generality of components may help in object recognition under transformations but reduces the accuracy when different objects have to be distinguished. For example, histogram based models are very robust to object transformations but lack specificity [13]. In the opposite side, components such as grayscale patches of objects that are view-specific cannot be used to recognize transformed versions of the same object [10], [12].

Hierarchical models inspired by the visual cortex have been used for transformation invariant object recognition. One of the earliest hierarchical models is the Neocognitron which was designed to perform position invariant object recognition [14]. Recent hierarchical models have been used to perform face and car recognition with high accuracy [10], [12]. These models are based on the idea of Perrett and Oram, who proposed that transformation invariance can be achieved by pooling over units that are tuned to different views of the same feature [15]. Convolutional neural networks are a subclass of hierarchical methods which have shown high performance in face and generic object recognition [16]-[17].

The models mentioned above lack the biologically plausibility for various underlying mechanisms; for example weight sharing has not been observed in brain neurons.
There are other models which have used more biologically plausible operations to perform object recognition and predict the properties of neurons in the visual cortex. A series of models of VisNet have been proposed that perform object recognition under translation, transformation and lighting variations [18]-[20]. Lissom is a set of hierarchically connected areas that utilize long range inhibitory and excitatory connections to produce the selectivity and recognition similar to neurons in the Inferotemporal cortex [21]. Serre and Poggio introduced a hierarchical model based on the HMax model that could perform object recognition in cluttered environments with high accuracy [22]. The main characteristic of their model was the alternating layers of simple and complex cells that provide recognition and invariance respectively.

In this paper, we propose a model for object recognition with emphasize on the issues of invariant object recognition and efficient feature coding. Compared to the Lissom and VisNet, our model uses efficient coding in each layer to extract the most informative features from images which results in higher recognition rates. In the next section, we describe the properties of the proposed model.

III. THE PROPOSED MODEL

Our model is composed of 3 layers corresponding to areas V1, V2 and V4 in the visual cortex. These areas are organized in a hierarchical manner as shown in Fig.1. Units in each layer receive inputs from the neurons in previous layer via bottom-up connections and in the same layer through horizontal connections. Units in the first layer receive bottom-up input from a small region of the input image. Bottom-up connections provide excitatory inputs to the units that determine their selectivity. Incoming signals through horizontal connections are inhibitory and facilitate the extraction of optimal features in images.

Neurons in the first layer are modeled by Gabor filters with different orientations:

\[
F(x, y) = \exp\left(-\frac{x_0^2 + y_0^2}{2\sigma^2}\right) \cos\left(\frac{2\pi x_0}{\lambda}\right)
\]

where \(x_0 = x \cos(\theta) + y \sin(\theta)\)

\[
y_0 = -x \sin(\theta) + y \cos(\theta)
\]

Here \(x\) and \(y\) are 2D coordinates of the filter in image, \(\gamma\) is the aspect ratio, \(\theta\) is the orientation, \(\sigma\) is the effective width and \(\lambda\) is the wavelength of the filter. We used a set of filters with 6 different orientations and only one spatial frequency (Fig.2).

For neurons in areas V2 and V4, bottom up activities are calculated as a weighted sum of responses of neurons in their receptive field from the previous layer:

\[
y_{i}^{\text{bup}} = \sum_{j \in RF_i} w_{ij}^{\text{bup}} x_j
\]

TABLE I displays the parameters of the proposed model.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Dimensions</th>
<th>Bottom-up RF</th>
<th>Horizontal Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>128*128</td>
<td>7*7</td>
<td>21*21</td>
</tr>
<tr>
<td>V2</td>
<td>104*104</td>
<td>21*21</td>
<td>21*21</td>
</tr>
<tr>
<td>V4</td>
<td>80*80</td>
<td>41*41</td>
<td>21*21</td>
</tr>
</tbody>
</table>

A. Learning Selectivity Invariance with Trace Learning Rule

The size of receptive field of neurons along the ventral visual pathway gradually increases and their preferred stimuli become more complex [2]. According to different electrophysiological experiments, neurons in V2 and V4 exhibit invariance selectivity to different transformations of...
their preferred stimuli [6]-[7]. It has been suggested that complete invariance selectivity can be developed in IT neurons by partial invariance responses of neurons in the areas V2 and V4 [23]. A neural mechanism has been proposed to achieve response invariance in preferred stimuli transformations [24]. This mechanism is based on the observation that activated neurons retain their activity level for hundreds of milliseconds. Fig.3 describes this mechanism. The basic idea is that the persistent activity level of neuron aids in strengthening the connections between this neuron and neurons that represent the transformed version of their preferred stimuli (see legend of Fig.3).

![Image](image_url)

**Fig. 3.** The idea of Continuous Transformation. (a) Neuron shown in blue is activated by a set of input from green neurons and the weights of connections between blue and green neurons are established. (b) As the stimulus transforms, the set of active neurons in the lower layer changes. Because the blue neuron preserves its activity level for a short time, it can establish connections with these new activated green neurons that represent the transformed version of the stimulus.

Of special interest in this paper is the Trace rule proposed by Földiák to learn the invariance selectivity of neurons for spatial and temporal transformations of the objects [25]. This rule extends the simple Hebbian learning rule to include a trace of the previous responses of neuron to update its weights:

\[
\nabla w_{ij} = \lambda \overline{y}_j^t (x_i^t - w_{ij}^{-1})
\]

(3)

In this equation, \(\overline{y}_j^t\) is the trace of neuron \(j\) in the iteration \(t\), \(w_{ij}\) is the weight of connection from input \(i\) to neuron \(j\), \(x_i^t\) is the input in iteration \(t\), and \(\lambda\) is the learning rate. The term \(-\lambda \overline{y}_j^t w_{ij}^{-1}\) is added to (3) to avoid unlimited increases in weights of connections. Trace value for a neuron is calculated using (4):

\[
\overline{y}_j^t = \eta \overline{y}_j^{t-1} + (1 - \eta) y_j^t
\]

(4)

where \(\eta\) is the trace constant which is selected in the interval [0,1] and \(y_j^t\) is the output of neuron \(j\) in iteration \(t\).

The trace learning rule is applied to develop the selectivity of neurons on their bottom-up input. The activity of each neuron induced from bottom-up connections is calculated as a weighted sum of activities of neurons in its receptive field from the previous layer. We used a subset of images from the coil100 dataset [33] to train the connectivity between layers of the model. Coil100 dataset contains images of objects from different view-points and is therefore appropriate to train the invariance selectivity of neurons in the model.

**B. Redundancy Reduction**

For an object recognition system to achieve high rates of recognition, efficiency of the features extracted from images is very important. A model that only tries to achieve invariance responses cannot result in high recognition rates for different objects. One must use features with high information content that can discriminate between objects. The redundancy reduction becomes an important issue when dealing with different objects from different categories. In the proposed model, we used a redundancy reduction mechanism that is based on a statistical model of the natural images. This mechanism is used in different layers to provide a set of globally optimal features for object recognition.

Primary efforts to reduce the redundancy in natural images revealed that linear filters like Gabor or Wavelet are optimized feature detectors as they are uncorrelated [26]-[27]. However dependencies in natural images are nonlinear and one cannot achieve independence with a linear transformation. Studies on the statistics of natural images discovered the nonlinear variance dependency between responses of linear filters in natural images [28]-[29]. Wainwright and Simoncelli proposed the Gaussian scale mixtures model to provide an explanation for variance dependency [30]. Schwartz and Simoncelli used divisive normalization to produce the independent responses over a set of patches selected from natural images [31]. This model was extended to a hierarchical architecture and simulated the properties of neurons in higher order visual area V2 [32]. We adapt this model in the proposed hierarchy to extract the most efficient features from natural images.

Neurons in each layer of the proposed model are connected to a set of neighboring neurons with horizontal connections. The weights of these connections are learned to predict the variance of responses of each neuron:

\[
\var(L_{\overline{x}} | L_{\overline{y}}, \overline{y} \in C_{\overline{x}}) = \sum_{\overline{y} \in C_{\overline{x}}} w_{\overline{xy}} L_{\overline{y}}^2 + \sigma_{\overline{x}}^2
\]

(5)

where \(L_{\overline{x}}\) and \(L_{\overline{y}}\) are responses of neurons \(\overline{x}\) and \(\overline{y}\) respectively and \(w_{\overline{xy}}\) is the weight of connection between them, \(C_{\overline{x}}\) is the neighboring region around \(\overline{x}\) and \(\sigma_{\overline{x}}^2\) is the part of variance of neuron \(\overline{x}\) that is independent of other neurons. Response of the neuron is then divided to its variance to remove the variance dependency. This mechanism is repeated in different layers of the model to produce neuronal responses that are globally independent:

\[
R_{\overline{x}} = \sum_{\overline{y} \in C_{\overline{x}}} w_{\overline{xy}} L_{\overline{y}}^2 + \sigma_{\overline{x}}^2
\]

(6)
the result of (6) is the independent response \( R_x \) of filter \( x \).

The weights of horizontal connections are learned in a training procedure after the bottom-up connections to the layer has been established. The goal is to predict the variance of responses of each neuron using responses of its neighboring neurons. An unbiased estimate for variance of responses of a neuron is:

\[
\bar{\sigma}_i^2 = \frac{1}{N-1} \sum_{t=1}^{N} \left( s_i^t - \bar{s}_i^t \right)^2
\]

\[\text{(7)}\]

An iterative form for the above equations is:

\[
\bar{s}_i^t = \left( \bar{s}_i^{t-1} \times (t-1) + L_i^t \right) / t
\]

\[
\left( \bar{\sigma}_i^2 \right)^2 = \left( \left( \bar{s}_i^{t-1} \right)^2 \times (t-2) + \left( L_i^t - \bar{s}_i^{t-1} \right)^2 \right) / (t-1)
\]

\[\text{(8)}\]

We used gradient descent to minimize MSE between variance estimates using (5) and (8). The update rules for horizontal connections are as follows:

\[
E^2 = \frac{1}{N} \sum_{i=1}^{N} \left( \text{Var}(L_i^t) - \left( \bar{\sigma}_i^2 \right)^2 \right)^2
\]

\[
\frac{\partial E^2}{\partial \sigma_i^2} = \frac{\partial}{\partial \sigma_i^2} \left( \left( \text{Var}(L_i^t) - \left( \bar{\sigma}_i^2 \right)^2 \right)^2 \right) = \sum_{i=1}^{N} \left( \text{Var}(L_i^t) - \left( \bar{\sigma}_i^2 \right)^2 \right) \times \left( L_i^t \right)^2
\]

\[\text{(9)}\]

\[
w_i^t = w_i^{t-1} - \eta \frac{\partial E^2}{\partial \sigma_i^2} = w_i^{t-1} - \eta \times \left( \text{Var}(L_i^t) - \left( \bar{\sigma}_i^2 \right)^2 \right) \times \left( L_i^t \right)
\]

\[\text{(10)}\]

**IV. EXPERIMENTAL RESULTS**

We performed a set of experiments to evaluate the performance of the proposed model for object recognition in general, under transformations and with different backgrounds. We used the coil100 dataset that contains images of different objects under different viewpoints to measure the generalization of the proposed model. The results are shown in Table.II for objects that were not used in training the network. It can be seen that the proposed model is superior to the other models reported in [21].

Images in the coil100 dataset do not challenge the ability of proposed model to recognize objects in different lighting conditions or background texture. Therefore, we generated a set of images from 3 objects (Fig.4) in different distances, viewpoints and backgrounds to examine the model. Samples for one object are shown in fig.5. For each object, we used two third of images for training and one third for test. In the first experiment, images for training were selected from all backgrounds. Results are shown in the second column of the table.III. In the second experiment, we used objects with two backgrounds for training and the third background was only used for test. Results are shown in third column of the table.III. It can be seen that the performance of the model is on average above 95%.

**V. CONCLUSION**

In this paper, we proposed a hierarchical model for object recognition based on well known properties of neurons in the primate visual cortex. Two key characteristics of the proposed model are the invariance of responses of its neurons with respect to object transformations and extraction of efficient features from images that enhance the object recognition accuracy. We used the trace rule to develop neurons with invariance responses to object transformation. In order to increase the accuracy of the recognition, we used a model of redundancy reduction previously used to develop neurons with similar selectivities to neurons in the visual cortex. We examined the performance of the model with a set of images of objects from different view-points and backgrounds. On average, the recognition rate was higher than 95% on coil100 and custom images.

Previous studies on redundancy reduction tried to develop neuronal selectivities similar to that of the primary visual cortex. In a recent study, redundancy reduction was extended to simulate the selectivity of extrastriate neuronal selectivities. In this paper, we proposed a hierarchical architecture to perform object recognition with features similar to selectivities of neurons in the V2 and V4. Neurons in these areas exhibit some degree of invariance response to stimuli transformations. Previous studies reported a high recognition rate with features inspired from visual cortex. In this paper, we provided a more biologically plausible mechanism for developing neurons with properties similar to the visual cortex neurons.
Fig. 5. Samples of images of object 1 from different viewpoints, distances and backgrounds.

<table>
<thead>
<tr>
<th>OBJECT</th>
<th>TRAIN AND TEST WITH DIFFERENT BACKGROUNDS</th>
<th>TRAIN WITH TWO BACKGROUNDS AND TEST WITH THE THIRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT1(DOLL)</td>
<td>0.957</td>
<td>0.956</td>
</tr>
<tr>
<td>OBJECT2(SHAMPOO)</td>
<td>0.916</td>
<td>0.975</td>
</tr>
<tr>
<td>OBJECT3(COFFEE MATE)</td>
<td>0.975</td>
<td>0.963</td>
</tr>
</tbody>
</table>

REFERENCES