V4 Neural Network Model for Visual Saliency and Discriminative Local Representation of Shapes

Hui Wei and Zheng Dong
Department of Computer Science, Laboratory of Cognitive Model and Algorithm, Fudan University
Shanghai Key Laboratory of Data Science, Fudan University
Shanghai, China
weihui@fudan.edu.cn

***Abstract*—Visual area V4 lies in the middle of the ventral visual pathway in the primate brain. It is an intermediate stage in the visual processing for object discrimination. It plays an important role in the neural mechanism of visual attention and shape recognition. V4 neurons exhibit selectivity for salient features of contour conformation. In this paper, we propose a novel model of V4 neurons based on a multilayer neural network inspired by recent studies on V4. Its low-level layers consist of computational units simulating simple cells and complex cells in the primary visual cortex. These layers extract preliminary visual features including edges and orientations. The V4 computational units calculate the entropy of the extracted features as a measure of visual saliency. The salient features are then selected and encoded with a layer of Restricted Boltzmann Machine to generate an intermediate representation of object shapes. The model was evaluated in shape distinction, handwritten digits classification, feature detection, and feature matching experiments. The results demonstrate that this model generates discriminative local representation of object shapes. It provides clues to understand the high level representation of visual stimuli in the brain.**

I. Introduction

Understanding the content of images has always been a difficult task in image analysis due to the well known semantic gap [1] between low-level representation of images and the highly abstracted semantic content contained in the images. However, biological brains accomplish this task accurately and effortlessly. It is an attractive goal to understand the neural mechanism of vision in the brain and simulate this mechanism with electronic computers. Neuroscience studies in the past decades have provided us with the opportunity to understand the neural processes of visual perception [2].

Visual stimuli captured by the eye are transformed into neural impulses in the retina and further processed by the visual cortex for high level cognitive tasks. The visual cortex plays an important role in filling the gap between visual stimuli and the implied semantic information. Neural impulses travel through the visual pathways in the visual cortex and finally contribute to a unified percept of the object of interest. Primate brains possess two distinct visual pathways [3], [4]. The dorsal pathway is involved with processing the object's spatial location relevant to the viewer. The ventral pathway is involved with object discrimination and recognition. In this paper, we concentrate on the latter, the ventral pathway, and visual area V4 in particular.

V4 lies in the middle of the ventral pathway (Fig. 1). Lower levels of the pathway (visual areas V1 and V2) extract preliminary visual features including edges and orientations [5], [6]. Higher levels of the pathway (inferior temporal cortex) exhibit selectivity to complex objects like faces and body parts [7], [8]. As an intermediate stage, V4 plays a crucial role in transforming low-level orientation signals into complex object representations.

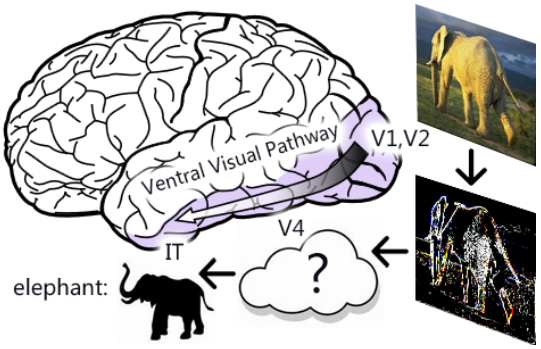


Fig. 1. V4 is an intermediate stage in visual recognition.

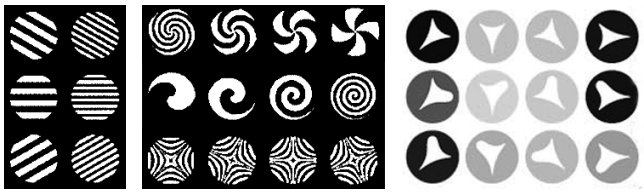
In the following section, we briefly introduce the neurobiological studies on V4 by which our model is inspired. We also discuss several previous work on the computational models of V4 and their limitations. In section III, we describe our model in details. The proposed model is a multilayer neural network. It extracts low-level orientation features from images and measures the visual saliency of the features. The salient features are further encoded into discriminative local representation of object shapes. In section IV, the model is evaluated in a series of experiments. The conclusion is summarized in section V.

II. Related Work

A. Shape Selectivity of V4 Neurons

Neurobiological studies on V4 have not produced a unified model of its function or circuitry. V4 neurons are known to be selective for color, shape, depth and even motion [9]. In this paper, we focus on the V4 selectivity for shapes. Early experiments examined the selectivity of cells in V4 with classical stimuli including bars and sinusoidal gratings (Fig. 2a) [10]. Similar to earlier processing stages, some V4 neurons are tuned for orientation and spatial frequency of edges and linear sinusoidal gratings. However, the majority of V4 neurons are sensitive to more complex shape properties. Later experiments demonstrated that V4 neurons display a clear bias in

their responses in favor of non-Cartesian gratings (Fig. 2b) and they show a significant degree of invariance in their selectivity across changes in stimulus position [11]. More recent experiments showed that V4 neurons can be strongly selective for curvature of contours and angular position of acute curvatures [12], [13]. Fig. 2c shows a response map of a V4 neuron (reproduced from [13]). The white shapes in Fig. 2c is presented in the receptive field of this V4 neuron with equally dark background. The gray scale in the response map indicates the strength of the neuronal response. Darker background indicates that the neuronal response is stronger. This neuron is selectively tuned for acute convex border at the bottom left. The curvature and the angular position are both necessary conditions of the activation of this neuron. Neither rounded protrusions nor sharp curvatures towards directions other than the bottom left activate the neuron.



(a) (b) (c)

Fig. 2. Shapes to examine V4 selectivity. (a) Classical gratings. (b) Non-Cartesian gratings. V4 neurons prefer non-Cartesian gratings rather than classical gratings. (c) Response map of a V4 neuron which responds to a sharp convex curvature at the bottom left. Darker background indicates a stronger response.

B. Previous Models of V4

Several models have been proposed to explain the shape selectivity and invariance of V4 neurons. We briefly introduce the SRF model [14] and the HMAX model [15], [16].

The spectral receptive field (SRF) [14] describes properties of V4 receptive field in terms of the orientation and spatial frequency spectrum. The model is based on the fact that V4 neurons have large orientation and spatial frequency bandwidth. They respond selectively to stimuli such as contour conformations and non-Cartesian gratings, which generally consist of multiple orientations and spatial frequencies. The spectral model is also invariant to small changes in stimulus position and thus explains the invariance property of V4 response patterns. The model is powerful in describing the shape selectivity of V4 neurons. However, it does not explain the emergence of the selectivity. It does not either provide the neural computing process to achieve such spectral receptive field.

The HMAX model was first proposed in [15] as a generic model for object recognition in the visual cortex. It models the visual cortex into a hierarchical architecture consisting of cascaded linear filters and non-linear maximum pooling operations. It was then adopted as a model for V4 shape selectivity and invariance [16]. The training of the model is an NP-complete problem. The authors used a greedy algorithm to obtain approximated solutions but they did not provide any biological evidence for the algorithm.

The proposed model of V4 in this paper overcomes the limitation of the previous models. In the next section, we demonstrate that its architecture and function is analogous to those of the visual cortex. We also provide efficient training method for our model.

III. Neural Network Model of V4

The information processing in the visual cortex follows a hierarchical scheme. Our model employs a similar hierarchical structure. Fig. 3 shows the architecture of our model.

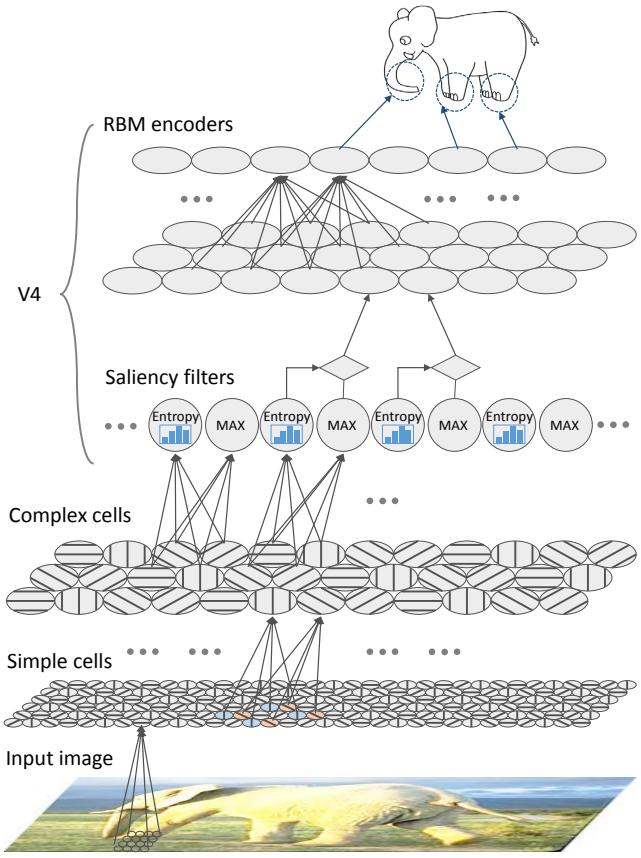


Fig. 3. Multilayer neural network model of V4.

In the feed-forward direction, the first two layers consist of simple cells and complex cells providing orientation features as input to V4 computation units. The layer of saliency filters measures visual saliency in terms of the entropy of features. Salient features are accurately located by maximizing neuronal response. The extracted features are then further encoded into local representation of object shapes.

A. Input Layers

According to the hierarchy of the ventral visual pathway, area V4 receives input from the lower levels including area V1 and V2. These areas have been well studied since 1960s by Hubel, Wiesel [5], [6] and succeeding researchers.

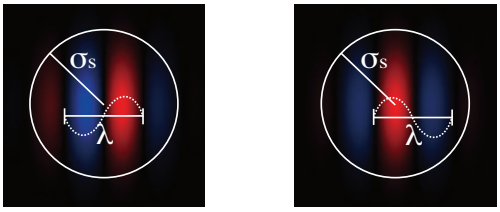
Neurons in V1 and V2 respond to local orientations. They fall into two categories, simple cells and complex cells. Simple cells respond primarily to oriented edges and gratings. Complex cells have larger receptive fields. A stimulus is effective

wherever it is placed in the complex receptive field, provided that the orientation is appropriate [5].

The receptive fields of simple cells can be understood as linear filters modeled as Gabor functions [17],

$$g_{\theta,\sigma_s}(x,y) = \exp\left(-\frac{x'^2+y'^2}{2\sigma_s^2}\right)\cos\left(2\pi\frac{x'}{\lambda}\right), \quad (1)$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. In the equation, θ represents the preferred orientation, σ_s approximates the radius of the receptive fields, and λ is the wavelength of the sinusoidal factor. λ controls the spatial frequency of the filter. It is twice the width of the central excitatory subregion of the receptive field (Fig. 4b). In this paper, it is taken according to the size of the simple receptive field. ($\lambda = 1.3\sigma_s$). We can have different Gabor functions by changing the phase offset of the sinusoidal factor. Fig. 4 shows two typical cases. They are equivalent with respect to extracting the orientation of edges. Equation 3 produces an even function (Fig.4b), which was used in this paper.



(a) Odd symmetry (b) Even symmetry

Fig. 4. Gabor functions with different phases.

In our model, the layer of simple cells operates on raw image input (Fig. 3). The output of simple cells with preferred orientation θ and scale σ_s is the following convolution passed through a transfer function ϕ,

$$S_{\theta,\sigma_s}(x,y) = \phi(I \otimes g_{\theta,\sigma_s}), \quad (2)$$

where I is an image and

$$\phi(u) = \begin{cases} u & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Complex cells are commonly thought as the squared summation of simple cells in phase quadrature [18]. In our model, complex cells are simplified as a linear summation of simple cells in different positions, weighted with a Gaussian function,

$$f_{\sigma_c}(x,y) = \frac{1}{2\pi\sigma_c}\exp\left(-\frac{x^2+y^2}{2\sigma_c^2}\right), \quad (4)$$

where σ_c is the scale of complex cells. Complex receptive field is usually 2 to 5 times larger than simple receptive field [19] and therefore in this paper, we have $\sigma_c = 2\sigma_s$.

In section IV, we show experimentally that complex cells provide sufficient information for the V4 model to form the selectivity for shapes.

B. Saliency Filters

V4 is an area of attentional modulation [9]. Visual attention involves selecting an interested region or selecting specific object features. Visual attention in V4 is influenced by both top-down feedback from higher levels in the visual pathway and bottom-up input from lower levels. We focus on the bottom-up influence. In a bottom-up process, V4 evaluates the saliency of the input from lower levels and focuses its attention automatically on the salient features.

We use entropy to measure the saliency of images [20]. In [20], it is assumed that salient regions have high complexity (and correspondingly high entropy).The entropy of features is used as a scale invariant measure. The salient region is selected at entropy peaks over scales. To avoid erroneously selecting noise or texture as salient regions, a measure of self-similarity is employed. Self-similar regions are filtered out.

In this paper, we use entropy in a different approach. V4 neurons encode fragments of object contour [13], [21]. They are selective for simple structures such as convex or concave curvature. Therefore, in our assumption, well ordered structures with low complexity are preferred in cognitive activities. We filter out regions with high complexity (or entropy).

The entropy is calculated according to the output of complex cells. Given a point (x,y), in the neighborhood of (x,y), a complex cell with preferred orientation θ has output value $C_\theta(x,y)$, We suppose that the probability of a complex cell being activated is proportional to the output value. The probability is thus defined as:

$$P(\theta) = \frac{1}{\sum_{\theta_i} C_{\theta_i}(x,y)} \cdot C_\theta(x,y). \quad (5)$$

The entropy of the complex cell activity in this neighborhood is

$$E = -\sum_\theta P(\theta)\log P(\theta). \quad (6)$$

Fig. 5 shows four image patches. For each image patch, the output values of complex cells with different preferred orientations (from 0° to 180°) are plotted in a bar chart. The charts show the distribution of complex cell activities. The entropy calculated accordingly indicates that patches composed of simple structures have non-uniform distributions and thus low entropy values. Therefore, we filter out regions with high entropy.

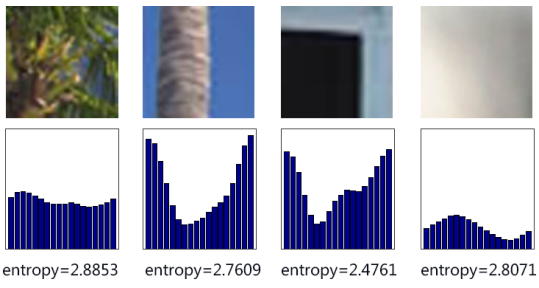


Fig. 5. Entropy of image patches. Bar charts in the second row show the output values of complex cells with different preferred orientations (from 0° to 180°).

In addition to entropy, local competition also plays a role in attentional selection [22]. With limited neural resources, only strong and competitive neuronal signals get transmitted and processed. In our model, the saliency layer finds local maximums of complex cell output and filters out those points with high entropy values or low activities (or output values). The algorithm is listed in Algorithm 1.

Algorithm 1 Saliency filter

procedure FINDSALIENTPOINT(image I, scale σ_c, threshold t_A, t_E)
  for each orientation θ **do**
    $C_\theta \leftarrow \phi(I \otimes g_\theta) \otimes f_{\sigma_c}$ ▷ complex cell output
  end for
  for each point (x, y) in image I **do**
    $C(x, y) \leftarrow \max_\theta C_\theta(x, y)$
    $E(x, y) \leftarrow$ entropy at point (x, y)
  end for
  Divide image I into patches of size $\sigma_c \times \sigma_c$
  for each patch p **do**
    $(\hat{x}, \hat{y}) \leftarrow \text{argmax}_{(x,y)\in p} C(x, y)$
    if $C(\hat{x}, \hat{y}) > t_A$ and $E(\hat{x}, \hat{y}) < t_E$ **then**
      Mark $(\hat{x}, \hat{y})$ as a salient point
    end if
  end for
end procedure

C. RBM Encoders

With saliency filters described in the previous subsection, we are able to focus on a limited number of salient points. The V4 computation units in our model encode the shape in the neighborhood of each salient point. The encoding is achieved with Restricted Boltzmann Machine (RBM).

RBM can learn a probability distribution over its set of inputs. It has been found efficient in training deep neural network [23]. The encoder layer in our model is part of a deep network. Therefore, we use the RBM as a training model for this encoder layer. The RBM is trained to encode shape features in the neighborhood of a salient point. Fig. 6 shows such an RBM encoder.

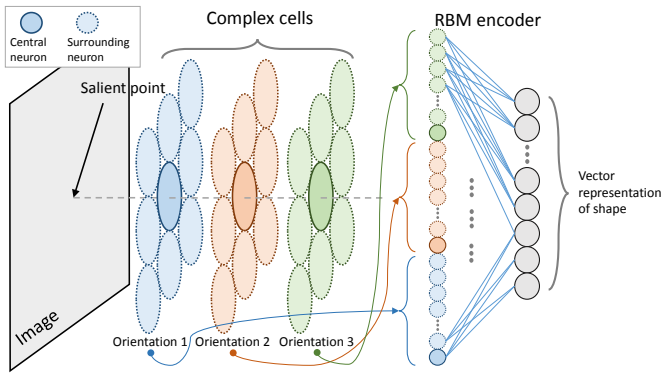


Fig. 6. RBM encoder for local shape feature.

Let (x, y) denote a salient point. For each preferred orientation, we select the complex cell of which the receptive field is centered at point (x, y), and its eight neighbors. Fig. 6 demonstrates an example of three preferred orientations. In this case, 9×3 complex cells are selected. The number of selected neurons depends on the number of orientations we choose. The selected complex cells are then used as the input layer (or visible layer) of the RBM. The representation of the shape in the neighborhood of (x, y) is formed in the output layer (or hidden layer) of the RBM.

RBM can be trained efficiently with a contrastive divergence learning algorithm [24]. Let w_{ij} be the weight of the connection from the i-th visible unit to the j-th hidden unit. In each learning iteration, the change in the weight is given by

$$\Delta w_{ij} = \epsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}} \right), \tag{7}$$

where ϵ is a learning rate, $\langle v_i h_j \rangle_{\text{data}}$ is the product of two units when the visible layer is given the data, and $\langle v_i h_j \rangle_{\text{recon}}$ is the product of two units when the visible layer is given the reconstruction ([24] for details on training RBM).

We use weight-decay [25] to reduce overfitting by adding an L2 penalty term, $\frac{1}{2}\lambda w^2$. The weight change is then given by

$$\Delta w_{ij} = \epsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}} \right) - \epsilon \lambda w_{ij}, \tag{8}$$

where λ is a weight-cost coefficient. We followed [25] for the choice of the coefficient λ and the learning rate ϵ in our experiments.

The output vector of the hidden layer forms a representation of local shape features. It can be used directly as input for classification tasks. We can also use the RBM connection weight matrix to initialize a multilayer neural network for supervised back-propagation training.

D. Model Parameters

The size of the receptive field increases along the hierarchy of the visual system. Lower levels have relatively small receptive fields while higher levels have larger receptive fields. The size of receptive field in our model follows the same scheme. The relationship is shown in Table I. V4 neurons receive afferent connections from 3×3 complex cells that have partially overlapped receptive fields. Therefore the radius of the V4 receptive field is not 3 times as long as that of the complex cell.

TABLE I. SIZE OF RECEPTIVE FIELD

Category	Radius of receptive field
Simple cells	σ_s
Complex cells	$\sigma_c = 2\sigma_s$
V4 RBM encoders	$2.5\sigma_c$

We used two schemes for the number of different preferred orientations of simple cells and complex cells. For black and white images of shapes and handwritten digits in our experiment, we used 4 orientations (from $0°$ to $135°$ in steps of $45°$). For gray scale images, we used 18 orientations (from $0°$ to $170°$ in steps of $10°$) in order to preserve more information.

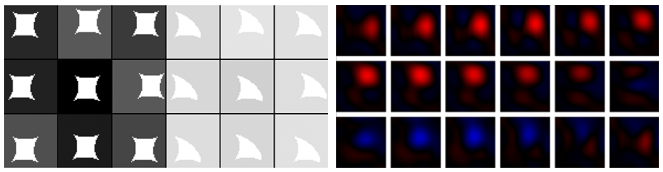
The output of complex cells was normalized to the $[0, 1]$ interval to serve as the input of RMB encoders. The saliency filters took a threshold $t_E = 2.8$ for the entropy and a threshold $t_A = 0.4$ for complex cell output value (or complex cell activity). These values were roughly the median values of natural images.

IV. Experiments

In this section, we demonstrate a series of experiments in which we evaluated our model.

A. Perceptron over Complex Cells

We have reviewed the shape selectivity of V4 neurons in section II. In the following experiment, We examined that the output of complex cells provides sufficient information for the emergence of neuronal response pattern of V4. A single perceptron was trained to distinguish between two shapes (Fig. 7a). The shapes were also used to examine the selectivity of V4 in [13]. The difference between the two shapes is that one has a sharp projection towards the top right.



(a) Response map (b) Weight matrices

Fig. 7. Response map and weight matrices of a perceptron that distinguishes two shapes. (a) Responses of the perceptron over 18 samples. The perceptron prefers the shape in the left half of the samples. It is insensitive to stimulus position. (b) The input weight of the perceptron. Each block shows the weight of connections from complex cells of a certain orientation. Complex cells of 18 different orientations provide input for this perceptron.

Since V4 neurons show a certain degree of invariance in their selectivity across changes in stimulus position, we moved the shapes randomly within the receptive field of the perceptron to generate samples for training and testing (Fig. 7a shows several samples). The samples were then used as input of the layers of simple cells and complex cells. The output of complex cells were passed to the perceptron. We had complex cells with different preferred orientations (from 0° to 170° in steps of 10°) and over different positions in the receptive field. Therefore, the input of the perceptron consisted of 18 matrices, each corresponding to the output of complex cells with a certain orientation. The input weight of the perceptron was thus also 18 matrices.

The trained perceptron exhibited a strong bias towards the shape with convex curvature towards the top right. It also showed a significant degree of invariance to the stimulus position. It is obvious that the selectivity is not formed from certain excitatory sub-regions or inhibitory sub-regions of the receptive field which was found of simple cells [5], [13]. The selectivity of the perceptron tallies with the selectivity of actual V4 neurons. The response map is shown in Fig. 7a. Darker background colors indicate stronger responses. The input weight of this perceptron is shown in Fig. 7b. Each block shows the weight from the complex cells of a certain preferred orientation. Red color denotes positive weight while blue color denotes negative weight.

This experiment demonstrates that complex cells provide sufficient information for V4 neurons to show selectivity observed in neurobiological experiments.

B. Shape Selectivity

In the following experiment, we trained our model to learn shapes. The training samples were collected from [11], [12], [13], including 4 categories of stimuli, i.e., sinusoidal gratings, non-Cartesian gratings, segmented curves, and closed shapes. The simple cells in our model took these sample images as input. The images were processed by simple cells and complex cells. Since the sample images were fitted into V4 receptive fields in neurobiological experiments, we adjusted the scale of the images to fit the size of a single V4 receptive field in our model too. Therefore the saliency filters were not necessary in this experiment. The output of complex cells were passed directly on to an RBM encoder.

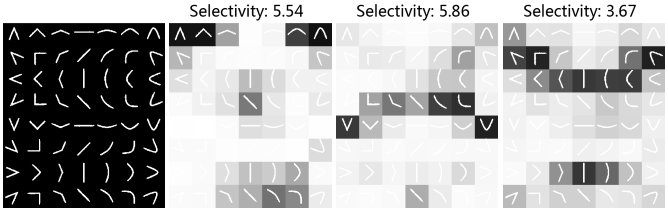
We used an RBM with 256 hidden units (RBM output neurons) in this experiment. In order to assess the selectivity of these units, we assigned each unit a selectivity index over a category of stimuli. The selectivity index is defined as the ratio of the maximal response to the average response over the stimuli of a certain category. Given an RBM output unit h and its output value h_i over a category of N stimuli for $i = 1, 2, \ldots, N$. The selectivity index $\mathcal{S}$ of h is given by

$$\mathcal{S}(h) = \frac{\max\{h_i | i = 1, 2, \ldots, N\}}{\sum_{i=1}^{N} h_i / N}. \tag{9}$$

In the following statistics, we assumed that a neuron has significant selectivity over a category of stimuli if the selectivity index is greater than 3.5. Among the 256 units, 171 units exhibited significant selectivity for different stimuli. 102 units showed selectivity for segmented curves (Fig. 8). 55 units showed strong bias towards non-Cartesian gratings (Fig. 9) while 50 units showed selectivity for classical sinusoidal gratings. 68 units exhibited selectivity for closed shapes (Fig. 10). The shape selectivity of our model is listed in TABLE II. The categories of selectivity are not mutually disjoint. A single unit may possess more than one kind of selectivity simultaneously.

TABLE II. Shape selectivity of our model

Selectivity	Number of Units	Percentage
Sinusoidal gratings	50	19.5%
Non-Cartesian gratings	55	21.5%
Segmented curves	102	39.8%
Closed shapes	68	26.6%

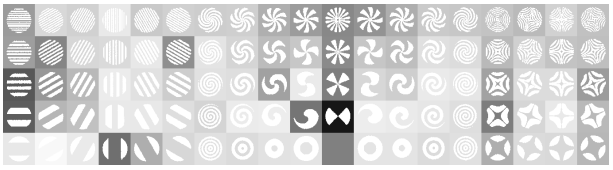


(a) Patterns (b) Response map (c) Response map (d) Response map

Fig. 8. RBM output neurons responding to convex curves towards certain directions. (a) Sample patterns of segmented curves. (b) Response map of a neuron tuned for curves projecting upwards. (c) Response map of a neuron tuned for curves projecting downwards. (d) Response map of a neuron tuned for curves projecting towards the top left.

Fig. 8a shows the sample patterns of segmented curves. Fig. 8b to Fig. 8d show the response maps of 3 RBM output

neurons. Darker background indicates stronger response. The selectivity indexes is shown at the top. It is shown that the three neurons exhibited strong bias towards curves rather than straight lines. The neurons were tuned for the orientation of the projection of the segmented curves. This is compliant with the selectivity of V4 neurons described in [12].



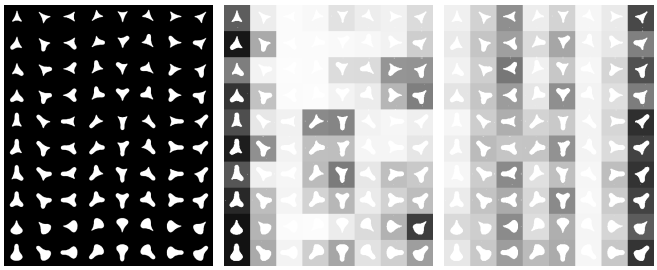
(a) Response map of a neuron selective for a pair of circular sectors (selectivity index = 3.92).

(b) Response map of a neuron selective for helix shapes (Selectivity index = 6.37).

Fig. 9. Response maps of RBM output neurons tuned for non-Cartesian gratings.

Fig. 9 shows the response maps of two RBM output neurons which are selective for non-Cartesian gratings. Both of the two neurons were not sensitive to sinusoidal gratings. The one in Fig. 9a was selective for a pair of circular sectors. The one in Fig. 9b preferred non-Cartesian gratings of helix shapes.

Fig. 10 shows the sample patterns of closed shapes and the response map of two neurons. The two neurons were selectively tuned for two kinds of shapes. They were sensitive to the angular position of the shapes.



(a) Patterns (b) Response map (c) Response map

Fig. 10. RBM output neurons responding to closed shapes.

This experiment demonstrates that our model can learn to distinguish the stimuli which V4 neurons are selectively tuned for.

C. Classifying Handwritten Digits

In the following experiment, we evaluated our model on the MINST database of handwritten digits [26]. We added another layer of artificial neurons over the RBM output to classify handwritten digits.

The images of digits are 28×28 pixels in size. We fitted the images into a single V4 receptive field thus the saliency filters were also bypassed in this experiment.

Since the images are black and white, we used complex cells with 4 preferred orientations, from 0° to 135° in steps of 45°. Thus the input layer of the RBM encoder consisted of 4×9 units. The output layer consisted of 128 units. We added another layer of 10 units to be the binary classifiers of the digits (from 0 to 9). These layers formed a $36 \times 128 \times 10$ feed-forward network. We trained the RBM encoder with the training set of data. The weight matrix of the RBM encoder was then used to initialize the weight between the first two layers of the network. The network was then fine-tuned with back-propagation (BP) algorithm [27]. We also trained the network directly with back-propagation algorithm for a comparison. Fig. 11 shows the training error. When the network was initialized with the RBM encoder of our model, the BP training began with a significantly smaller training error.

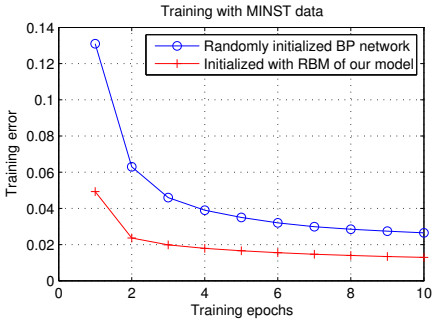


Fig. 11. Training error is smaller when using the RBM encoder to initialize the BP network.

The test error of our model is also competitive compared with other algorithms. A comparison is shown in Table III. The benchmark is provided in [26].

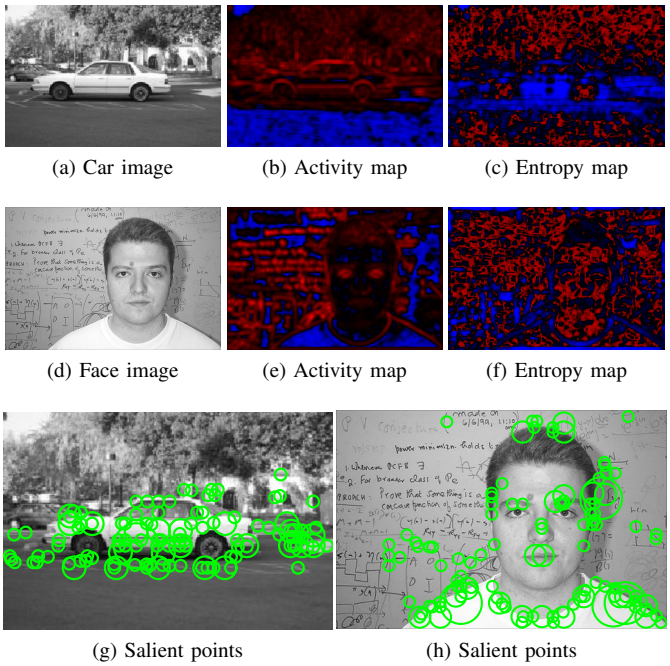
TABLE III. COMPARISON OF PERFORMANCE ON MINST DATABASE

Methods	Test error rate (%)
Our model	2.9
K-nearest neighbors	5.0
40 PCA + quadratic classifier	3.3
2-layer neural network	4.7
SVM, Gaussian kernel	1.4

D. Feature Detection

In the above experiments, the images contain very limited information. They are small enough to fit into the receptive field of V4 computation unit. When we deal with larger images, especially natural images, we have to find some regions of interest and focus the computation units over a limited number of interesting regions. A full scan over the whole image is not computationally economic. It also complicates succeeding processing for high-level tasks such as object recognition and scene understanding. The visual neural system takes a similar approach. As we have reviewed in previous sections, V4 is closely related with selective visual attention [9]. Research shows that V4 receptive field shrinks and shifts towards saccade target [28]. We simulated such visual attention with the saliency filters in our model. In the

following experiment, we evaluated the saliency filters with feature detection experiment.



(a) Car image (b) Activity map (c) Entropy map

(d) Face image (e) Activity map (f) Entropy map

(g) Salient points (h) Salient points

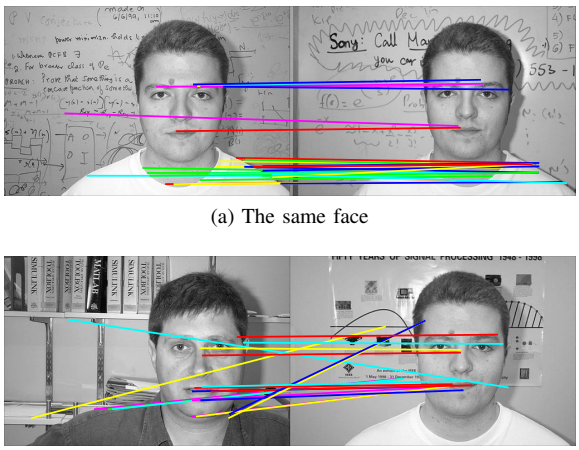
Fig. 12. Feature detection. (a) and (d) show the original images. (b) and (e) are the complex cell activity. Red color indicates strong activity and blue color indicates inhibition. (c) and (f) are the entropy map. Red color indicates high entropy value and blue color indicates low entropy value. (g) and (h) are the selected salient points at the scale of 4, 8, and 16 in terms of simple cell radius.

Fig. 12 shows the process of feature detection with the saliency filters in our model. Fig. 12b and Fig. 12d show the activity map of complex cells. The activities are high in the region of the trees behind the car (Fig. 12a) and the handwriting on the white board (Fig. 12c). Saliency filters filter out these regions because these regions have a comparatively high entropy. These areas consist of a large amount of disordered edges which result in a near uniform distribution of complex cell activity over preferred orientations and thus high entropy values. The saliency filters can find salient points at each given scale. Fig. 12g and Fig. 12h show the salient points at the scale of 4, 8, and 16 in terms of the simple cell radius (The size of the other units can be inferred with the relationship shown in Table I).

E. Feature Matching

Previous work on feature matching emphasizes on matching physical points on the same object under different views (e.g. [29]). The human vision has better generalization ability at feature matching in that we can match the same part of different objects of some certain category despite the minor differences in details. This ability is important for learning the semantic of objects. The following feature matching experiment demonstrates that our model exhibits this kind of generalization ability.

In this experiment, we trained our model with images randomly selected from the Caltech data set [30]. We took the output of the RBM encoders as the feature descriptors of the salient points. A match between points from two images was then established by finding pairs of features with minimal Euclidean distance. The matches with feature distances greater than a threshold value (1.25 in this experiment) were filtered out as non-matches. The feature matching result between face images is demonstrated in Fig. 13. Correct matches were established between different images of the same face. Different faces were also matched because they were visually similar.



(a) The same face

(b) Different faces

Fig. 13. Feature matching result between face images.

The comparison in Fig. 14 demonstrates the advantage of our model. We used SIFT feature [31] for comparison. Fig. 14a shows that our model produced correct matches between the model van in the left and the two real vans in the right. However, SIFT feature failed to produce correct matches (Fig. 14b). SIFT feature utilizes the direction of image gradient. It fails to capture the shape feature of an object when the object's color changes dramatically because the direction of image gradient changes with the color. Our model focuses on the shape feature and thus achieves a stable matching despite the change in color.

V. Conclusion

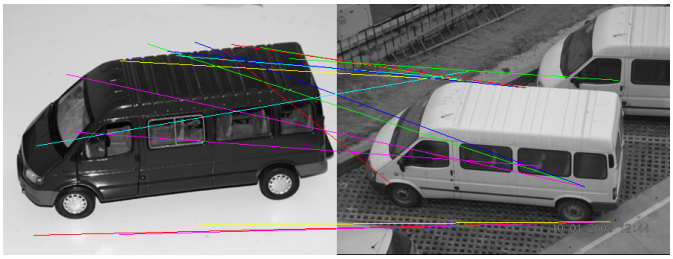
In this paper, we propose a model for the visual area V4. It is based on the neural mechanism of the ventral visual pathway. We focus on visual attention and shape selectivity in V4. V4 in our model is implemented as a multilayer neural network which selects salient points and encodes the shape feature in the neighborhood of the salient points. We demonstrate a variety of experiments in which the model was evaluated. The results show that the proposed model is consistent with the shape selectivity of area V4. It can find salient points in images and encode local shape feature into a discriminative representation.

Future work should involve further quantitative evaluations of our model. We will also investigate the application of our model in computer vision tasks such as object recognition and scene understanding.

Acknowledgment

This work was supported by the 973 Program (Project No. 2010CB327900), the NSFC project (Project No. 61375122,

(a) Our matches



(b) SIFT matches

Fig. 14. Comparison of feature matching between van images.

81373556), and the National "Twelfth Five-Year Plan" for Science and Technology (Project No. 2012BAI37B06).

REFERENCES

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] L. G. Ungerleider and A. H. Bell, "Uncovering the visual alphabet: advances in our understanding of object perception," *Vision research*, vol. 51, no. 7, pp. 782–799, 2011.

[3] G. Ettlinger, "object vision and spatial vision: The neuropsychological evidence for the distinction," *Cortex*, vol. 26, no. 3, pp. 319–341, 1990.

[4] S. R. Lehky and A. B. Sereno, "Comparison of shape encoding in primate dorsal and ventral visual pathways," *Journal of neurophysiology*, vol. 97, no. 1, pp. 307–319, 2007.

[5] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[6] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat." *Journal of neurophysiology*, 1965.

[7] C. Bruce, R. Desimone, and C. G. Gross, "Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque," *J Neurophysiol*, vol. 46, no. 2, pp. 369–384, 1981.

[8] A. H. Bell, F. Hadj-Bouziane, J. B. Frihauf, R. B. Tootell, and L. G. Ungerleider, "Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging," *Journal of neurophysiology*, vol. 101, no. 2, pp. 688–700, 2009.

[9] A. W. Roe, L. Chelazzi, C. E. Connor, B. R. Conway, I. Fujita, J. L. Gallant, H. Lu, and W. Vanduffel, "Toward a unified theory of visual area v4," *Neuron*, vol. 74, no. 1, pp. 12–29, 2012.

[10] R. Desimone and S. J. Schein, "Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form," *Journal of neurophysiology*, vol. 57, no. 3, pp. 835–868, 1987.

[11] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. Van Essen, "Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey," *Journal of neurophysiology*, vol. 76, no. 4, pp. 2718–2739, 1996.

[12] A. Pasupathy and C. E. Connor, "Responses to contour features in macaque area v4," *Journal of Neurophysiology*, vol. 82, no. 5, pp. 2490–2502, 1999.

[13] A. Pasupathy and C. E. Connor, "Shape representation in area v4: position-specific tuning for boundary conformation," *Journal of Neurophysiology*, vol. 86, no. 5, pp. 2505–2519, 2001.

[14] S. V. David, B. Y. Hayden, and J. L. Gallant, "Spectral receptive field properties explain shape selectivity in area v4," *Journal of neurophysiology*, vol. 96, no. 6, pp. 3492–3505, 2006.

[15] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[16] C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber, and T. Poggio, "A model of v4 shape selectivity and invariance," *Journal of Neurophysiology*, vol. 98, no. 3, pp. 1733–1750, 2007.

[17] D. Gabor, "Theory of communication. part 1: The analysis of information," *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, vol. 93, no. 26, pp. 429–441, 1946.

[18] D. J. Fleet, H. Wagner, and D. J. Heeger, "Neural encoding of binocular disparity: energy models, position shifts and phase shifts," *Vision research*, vol. 36, no. 12, pp. 1839–1857, 1996.

[19] J. Movshon, I. Thompson, and D. Tolhurst, "Receptive field organization of complex cells in the cat's striate cortex." *The Journal of physiology*, vol. 283, no. 1, pp. 79–99, 1978.

[20] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[21] A. Pasupathy and C. E. Connor, "Population coding of shape in area v4," *Nature neuroscience*, vol. 5, no. 12, pp. 1332–1338, 2002.

[22] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[23] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, 2010.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 1, p. 213, 2002.

[28] F. H. Hamker and M. Zirnsak, "V4 receptive field dynamics as predicted by a systems-level model of visual attention using feedback from the frontal eye field," *Neural Networks*, vol. 19, no. 9, pp. 1371–1382, 2006.

[29] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 43–57, 2011.

[30] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.