Article

Chinese Science Bulletin

Artificial Intelligence

October 2012 Vol.57 No.30: 3943–3952
doi: 10.1007/s11434-012-5402-x

A computational coding model for saliency detection in primary visual cortex

ZOU Qi*, WANG Zhe, LUO SiWei, HUANG YaPing & TIAN Mei

Department of Computer Science, Beijing Jiaotong University, Beijing 100044, China

Received April 5, 2012; accepted July 10, 2012

This study researches the coding model adaptive for information processing of the bottom-up attention mechanism. We constructed a coding model satisfying the neurobiological constraints of the primary visual cortex. By quantitatively changing the coding constraints, we carried out experiments on images used in cognitive psychology and natural image sets to compare the effects on the saliency detection performance. The experimental results statistically demonstrated that the encoding of invariant features and representation of overcomplete bases is advantageous to the bottom-up attention mechanism.

visual attention, coding in primary visual cortex, saliency detection, invariant feature, overcomplete bases

Citation: Zou Q, Wang Z, Luo S W, et al. A computational coding model for saliency detection in primary visual cortex. Chin Sci Bull, 2012, 57: 3943–3952, doi: 10.1007/s11434-012-5402-x

One of the fundamental problems of bottom-up attention is how the primary visual cortex encodes low-level features and forms a saliency map. Since the influential work of Treisman and Gelade [1], different coding models have emerged. Some are motivated by an imitation of information processing mechanisms in the primary visual cortex, to provide an efficient input to visual attention. For example, they simulate simple cells in V1 by Gabor filters, and implement multiscale processing by Gaussian pyramids [2–4]. Besides, the effects of overcomplete bases on encoding a bottom-up saliency map is of current interest [5,6]. While more and more neurobiological properties of the primary visual cortex have been accepted, the way they modulate the saliency detection remains unclear [7]. Models limited to simulation of simple cells cannot sufficiently satisfy neurobiological constraints [8]. How other mechanisms beyond simple cells affect bottom-up attention is worth deeply researching.

Existing computational models, which simulate coding mechanisms in the primary visual cortex, represent saliency according to the feature combination theory of Treisman and Gelade [1]. The first computational framework based on a neurobiological understanding of selective attention was proposed by Koch and Ullman [9]. Under this framework, the pre-attentive mechanism which extracts early visual features is often implemented by linear filtering using Gabor or other wavelets to simulate information processing in the primary visual cortex [3,10]. However, this simplified imitation may overlook the more complicated early vision mechanisms [2].

As we have known, the distribution of neurons in the primary visual cortex is hierarchical, showing specific topology [11]. The ability of invariant representation increases along a hierarchy [12]. Neurons in V1 expand the input from ganglion cells and adopt an overcomplete representation [6]. Therefore, as we construct coding models simulating the primary visual cortex we must consider these constraints to provide efficient input to bottom-up attention models.

The main contribution of this paper is to build a coding model, which satisfies the neurobiological constraints of the primary visual cortex, to provide efficient input for bottom-up attention. In this paper, two questions were addressed. First, does invariant representation in the primary visual cortex affect bottom-up attention? Second, what is the effect

*Corresponding author (email: qzou@bjtu.edu.cn)

of overcomplete representation on saliency detection? To our knowledge, such coding constraints have not been used in attention models. We constructed an attention model based on a coding model including these constraints. By quantitatively changing the coding constraints, we carried out experiments on images used in cognitive psychology and natural image sets to compare the effects on the saliency detection ability caused by different coding constraints.

1 Background

1.1 Computational models for bottom-up attention

Bottom-up attention models extract multi-dimensional features from an image and combine these features into a saliency map where the most salient object will be perceived.

In the feature extraction stage, computational models motivated by imitation of the primary visual cortex often use Gabor filters to extract orientation information at different scales. Properties of Gabor filters resemble simple cells' receptive fields and can provide input to the bottom-up saliency map. Similar methods also use Gaussian pyramids [13], Fourier transformation, or wavelets decomposition [10] to extract features similar to the responses of cells. One of representative models proposed by Itti et al. [3] adopted Gaussian pyramids to extract color, intensity, and orientation features at different levels. Grigorescu's model [14] simulated complex cells and nonclassical receptive field inhibition to detect salient contours.

Other researchers proposed powerful attention models. Poggio's group used the Bayes model and combined top-down features and spatial priors to generate a salience map [15,16]. Their results showed that using shape and spatial priors can improve saliency detection performance even in clutters and occlusions. Liu et al. [17] used a conditional random field to learn to detect a salient object. He combined multiple features including contrast, color and center-surround histogram. Our model differs from theirs in that we do not use any prior and do not learn any features from hand-labeled training set. Our model is a purely bottom-up attention model.

1.2 Coding models simulating information processing mechanisms in primary visual cortex

Traditional sparse coding models simulating V1 area estimate bases similar to receptive fields of simple cells by learning statistical properties of natural images [19]. This cannot account for the whole picture of early vision. A new area of research has emerged which designs nonlinear methods to capture topological relationships. Hyvärinen et al. [20] proposed the independent subspace analysis (ISA) and the topographic independent component analysis (TICA) [11]. Both models can extract the phase invariant and shift invariant features similar to the responses of complex cells. Wang et al. [21] developed a more computationally efficient model based on pairwise cumulant based methods for independent component analysis (PCICA). It captured the topological relationships by the pairwise cumulant and obtained invariant features. It converges faster than ISA and TICA and can be extended to overcomplete bases set.

Overcomplete representation is another important property in the primary visual cortex. Despite our recognition of its usefulness in early vision, we have not fully understood its role in forming a saliency map [6]. Several coding models considering overcomplete basis sets have been proposed, such as TICA and PCICA. PCICA is successfully used in object recognition. Recently, some saliency detection models using matrix decomposition learned overcomplete bases from color images [5]. Although it uses sparse coding and overcomplete bases, it is less biologically motivated and more mathematically implemented.

2 Saliency detection model based on coding in primary visual cortex

Given the neurobiological constraints on the information processing in the primary visual cortex, we constructed a coding model with overcomplete topological bases to extract features and form a saliency map. It comprises three steps: first, overcomplete topological bases are learned from training images. A test image is filtered by the basis set to produce initial features. Second, pooling the outputs of topological bases in neighboring regions obtains invariant features. Third, after suppression modeling the function of lateral connections between neurons, the features are combined to form a saliency map, and the focus-of-attention shifting sequence is determined accordingly. We describe the workflow in Figure 1.

2.1 Primitive features extraction by overcomplete topological bases

The model learns a set of overcomplete topological basis vectors $\{\phi_i\}$ from natural images by PCICA algorithm (for details refer to [21]), and computes their responses to an image patch $I(x,y)$ according to the following formula,

$$SF_i(x,y) = \phi_i^{\mathrm{T}} \left\| \phi_i \right\|^{-1} \hat{I}(x,y), \tag{1}$$

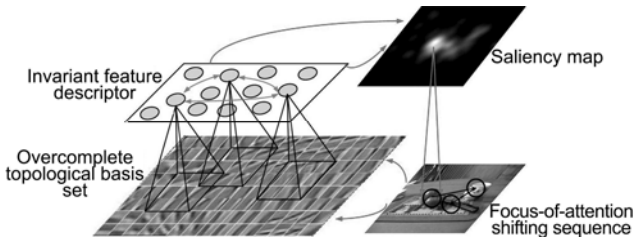

Figure 1 Workflow chart of our model.

where $\hat{I}(x, y)$ denotes a preprocessed image patch with the center located at (x, y). A test image is divided into patches with the same dimension to a basis. For example, if the size of ϕ_i is 16×16 pixels, the size of a patch is also 16×16 pixels. The preprocessing is that all the patches are whitened and their dimensions are decreased by principle component analysis.

As each basis responds optimally to specific frequency, phase, and orientation, local primitive features similar to simple cell responses are encoded by eq. (1). A set of 392 overcomplete bases learned from natural images are listed in Figure 2. In the preprocessing, all the 16×16 patches are whitened and their dimensions are decreased from 256 to 196. As the number of bases is two times the dimension of a basis, it is designated as two times overcomplete bases. The bases learned from natural images show clear topography. Orientations, frequencies, and locations of all the filters smoothly vary, forming a globally and topographically ordered array. Properties of filters in neighborhoods are similar.

2.2 Invariant features representation

Invariant features are obtained by organizing the responses of topological bases in the same neighborhood with the pooling operations. For the two times overcomplete bases as mentioned in section 2.1, we set the size of neighborhood to be 5×5 (an example is denoted by a box in Figure 2), and two adjacent neighborhoods overlap by two bases in both rows and columns. In this way, invariant feature descriptors Ω_j are obtained and each of them consists of 25 bases. Bases in the same descriptor correlate strongly, while bases in different descriptors usually correlate weakly.

The pooling cannot avoid some bases with strong differences being grouped into the same neighborhood, so a further refinement is needed. In every descriptor, each basis ϕ_i is compared with the basis located at the center ϕ_c by considering nonlinear correlations

$$\rho_{\phi_i\phi_c} = \operatorname{cov}\left(SF_i, SF_c\right)\left(\operatorname{var}\left(SF_i^2\right)\operatorname{var}\left(SF_c^2\right)\right)^{-1/2}, \qquad (2)$$

where cov(,) and var( ) denote covariance function and variance functions, respectively. SF_i was computed by using the eq. (1). In a descriptor, the bases correlating weakly to the center basis, indicated by $\rho_{\phi_i\phi_c}$ under some threshold, are removed. After this processing, the filters in the same descriptor have similar properties. Figure 3 shows nine examples of invariant feature descriptors obtained by pooling and then refining with a correlation threshold of 0.1.

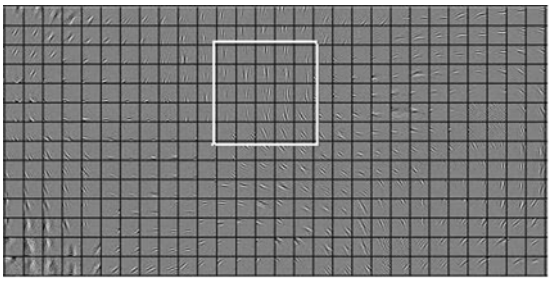

Figure 2 Overcomplete bases obtained by PCICA. A clear topography emerges from this map. An example showing the filters in neighborhood slowly change their properties as marked within a square.

After the invariant feature descriptors are determined, the responses $SF_i(x,y)$ of all filters ϕ_i belonging to the same descriptor Ω_j are pooled,

$$CF_j\left(x, y\right) = \sqrt{\sum_{\phi_i} SF_i^2\left(x, y\right),\ \phi_i \in \Omega_j} \quad \phi_i \in \Omega_j, \qquad (3)$$

where $CF_j(x, y)$ is the output of an invariant feature descriptor. As the orientations and frequencies of filters belonging to the same descriptor vary smoothly, we can obtain invariance by this pooling operation. It is biologically plausible that a bank of receptive fields at nearby locations on one level are organized to provide input to a receptive field on a higher level [22]. By pooling, the size of a receptive field on a high level is enlarged compared with the one on a low level, and its robustness to changes is increased as well. Simple cells and complex cells in the V1 area are an example. Receptive fields of simple cells overlap with each other, and those with similar properties pool to form receptive fields of complex cells.

2.3 Saliency maps formation

After an image is encoded by invariant feature descriptors, the parts which differ the most in a feature map are selected as candidates for salient objects. This is known as a preattentive process and it produces competitive results called conspicuous maps. Motivated by neurobiology, we obtained a conspicuous map by modeling suppression between neurons with a difference-of-Gaussian (DoG) operator, and then obtained the final saliency map by combining the conspicuous maps. Electrophysiological experiments show that neurons in the primary visual cortex of the macaque monkey modulate nearby neuronal responses by surround suppression. In an orientation feature channel, approximately 2/3 neurons produce the strongest suppression on neurons whose sensitive orientations are orthogonal to themselves, and produce little suppression on neurons whose sensitive orientations are parallel to themselves [23]. This suppression is strong within a certain range, and decays as the distance becomes smaller or larger [24]. It is called nonclassic receptive field suppression since the distance is beyond the range of classic receptive fields. The suppression satisfying the above conditions can be modeled by a DoG function

$$\operatorname{DoG}\left(x, y\right) = \frac{1}{2\pi\sigma_{\mathrm{ex}}^2}\mathrm{e}^{-\left(x^2+y^2\right)/2\sigma_{\mathrm{er}}^2} - \frac{1}{2\pi\sigma_{\mathrm{inh}}^2}\mathrm{e}^{-\left(x^2+y^2\right)/2\sigma_{\mathrm{inh}}^2}, \qquad (4)$$

where σ_{ex} and σ_{inh} indicate excitation and inhibition bandwidth.

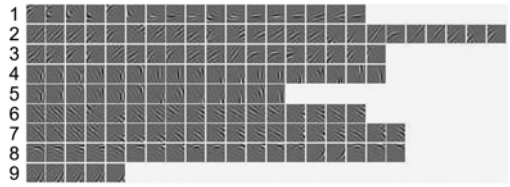

Figure 3 Some examples of invariant feature descriptors obtained by pooling and then refining. Each row lists a descriptor Ω_j in which orientations and frequencies are similar but phases are different.

This kind of suppression acts on the output of invariant feature descriptors, forming a conspicuous map

$$s_j(x,y) = \left| CF_j(x,y) + \alpha \left(CF_j \otimes \mathrm{DoG} \right)(x,y) \right|_{>0}, \qquad (5)$$

where α is a coefficient adjusting the strength of suppression, $\otimes$ denotes convolution, $|\ |_{>0}$ remains unchanged if inputting is positive, and outputs zero if inputting a negative or zero.

Finally, conspicuous maps are integrated into a saliency map by certain combination strategy. Considering the large number of conspicuous maps, direct summation or summation after normalization is not robust to noises and leads to a lot of local maxima. We adopted a combination strategy of iteration [25]. After iterating each conspicuous map by a DoG operator, we combined the conspicuous maps by summing them into a saliency map, which had sparse maxima.

3 Experiments

To research different coding model performances on saliency detection, we compared the results on two kinds of dataset: testing images widely used in cognitive psychological experiments on visual attention, and natural images. The experimental environment is MATLAB, running under Intel Core i5 2.66 GHz CPU.

3.1 Effects of invariant features on saliency detection

We used the gray images dataset (ww.cis.hut.fi/projects/ica/data/images) which is the standard dataset used in ICA and sparse coding models. We obtained 50000, 16×16 pixel patches from the training dataset and preprocessed them with whitening and dimension reduction. To compare the performances of saliency detection models with and without invariant features coding, we did the following:

(1) We obtained the overcomplete topological bases by PCICA, as shown in Figure 2, to extract primitive features, including localized, oriented, and bandpass edges.

(2) The primitive features, when in step (1), were pooled and refined to form invariant features descriptors. For the model without invariant features coding, this pooling was not needed, since the saliency map was obtained by directly performing DoG suppression on feature maps from the first step and from a subsequent combination.

(3) For the model with invariant features coding, the saliency map was obtained by performing DoG suppression on invariant features when in step (2), and a subsequent combination.

The first experiment is similar to the "visual search" tasks designed by Treisman and Gelade [1]. According to their results, stimuli that differ from homogeneous surrounding stimuli in orientations easily pop out. The tasks will be more difficult if distractors become less similar [6]. We designed dissimilar distractors whose orientations varied within the ranges from [−5°,5°] to [−45°,45°], to test performances of saliency detection. Sensitivities of the attention models on dissimilarities in distractors are shown in Figure 4.

We generated synthetic images as follows. Object positions and orientations were randomly determined. Distractors orientations were orthogonal to those of objects and were disturbed by orientation noises (in $[-\Delta\theta,\Delta\theta]$ with uniform probability). The bigger $\Delta\theta$, the more dissimilar are the distractors. Correct detection rates of attention models with and without invariant features coding drops with increased dissimilarities among distractors. This is consistent with the conclusion in psychological experiments. However, the model with invariant features coding is more robust to dissimilarities among distractors. When $\Delta\theta$ reaches 35°, it can still correctly detect objects with a rate of 0.75. This indicates that invariant features coding provides attention models relatively robust to disturbances.

We further tested the models' performances in detection of global salient structures. In contrast to visual search tasks, where salient objects are local points, the global saliency is defined by Gestalt psychologists in "figure-background segregation" tasks [18]. When local primitives form a structure satisfying certain perceptual organization rules such as proximity or good continuity, they will be perceived as a figure from the background. The S curve, an illusory contour and a noisy version were taken as testing images. Results are shown in Figure 5.

As indicated in Figure 5, objects as solid curves are perceived clearly. When objects are illusory contours, the attention model with invariant features coding can detect continuous contours. As parameters vary within the range mentioned below the figure, detection results are always continuous contours, while the attention model without invariant features coding detects discrete end points. When the object is a noisy s curve, differences in saliency detection results between the two attention models are greater. In the model with invariant coding, segments forming the s curve can be detected and most segments forming background can be suppressed. For the model without invariant coding, segments in the background are more salient than those in the s curve.

When objects are illusory contours like the middle image

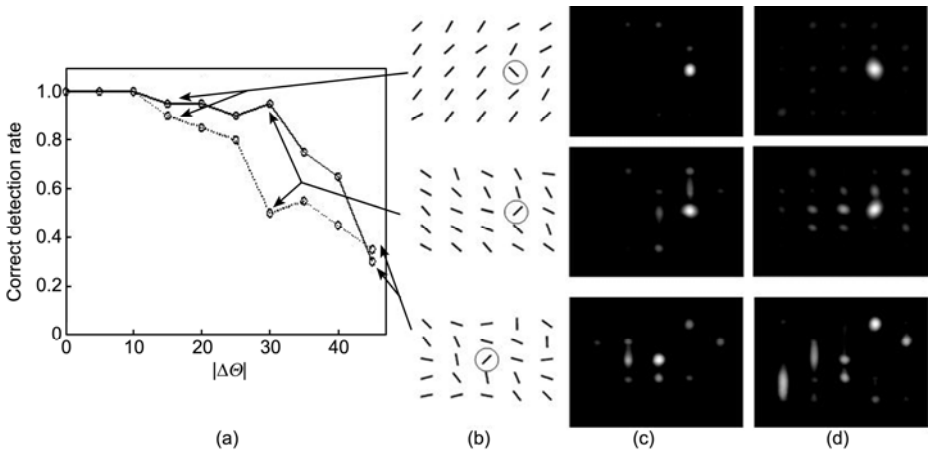

Figure 4 Saliency detection by two models in visual search where stimuli differ in orientations. (a) Statistics of correct detection rate *vs*. varying range of distractors' orientation $|\Delta\theta|$. Results of attention models with and without invariant features coding are denoted by red and blue lines, respectively. $|\Delta\theta|$ is from 0° to 45° at 5° intervals; (b) from up to bottom: examples of testing images whose $|\Delta\theta|$ are 15°, 30°, and 45°, respectively. The objects marked by red circles; (c) saliency maps of attention model with invariant features coding on testing images in (b); (d) corresponding saliency maps of attention model without invariant features coding.

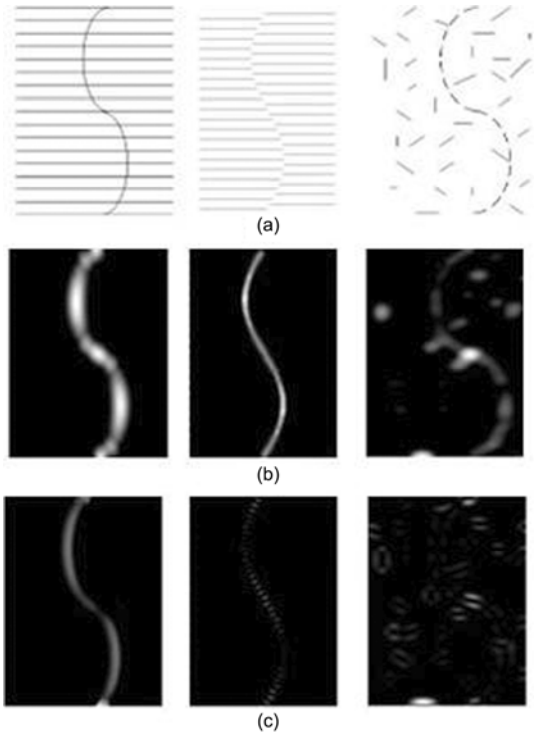

Figure 5 Saliency detection in figure-ground segregation as assessed by two models. (a) Source images; (b) saliency maps of attention model with invariant features coding; (c) saliency maps of attention model without invariant features coding. Results are relatively insensitive to parameters. σ_{ex} is 2%–5% of the image size (the bigger between image width and height). σ_{inh} is 4 to 10 times σ_{ex}, $\alpha \in [3,4]$.

in Figure 5, our model has end-inhibition groups thanks to the invariant representation, so that saliency detection result shown as discrete end points can be strengthened and connected into a continuous contour. Some of the complex cells in V1 are sensitive to the end of a line or edge. This property of responding strongly to either an edge, a bar or a slit which ends within the receptive fields is called end-inhibition [27] (Figure 6(a)). Such end-inhibition receptive fields can be found from invariant feature descriptors learned in our model. The sixth line in Figure 3 is an end-inhibition group, from which we select several typical pairs and list them in Figure 6(b).

Without invariant representation, responses to end points are occasional and scattered, so it is difficult to form a continuous structure. With invariant representation by pooling receptive fields of similar selectivity, responses to end points are concentrated in a group. The whole group's consistent and synchronous responses enhance saliency of the target so that discrete end points are connected into a salient continuous contour as shown in Figure 5(b).

When the object is a noisy s curve like the right image in Figure 5, invariant representation enhances the saliency of the s contour. A number of neurophysiological studies [28–30] have shown that invariant representation facilitates contour completion. It is observed that when an oriented stimulus appears within the receptive field of a neuron, and a second collinear or cocircular stimulus appears within the receptive field of another neuron which has similar orientation selectivity to the first neuron, the two neurons will increase the response rate of each other. As a result, they have an amplification

Figure 6 End-inhibition receptive fields. (a) An ideal example of end-inhibition receptive field used in neurophysiological experiments; (b) four pairs of end-inhibition receptive fields in an invariant feature descriptor learned by our model.

effect on the contour consisting of collinear or cocircular elements.Just as the image in Figure 5, the s contour consisting of aligned cocircular segments is amplified by the increasing responses from a group of neurons in an invariant descriptor, so the s contour pops out of clutter.

To compare saliency detection performances between the two models under different signal-to-noise ratios (SNR), we designed figure-ground segregation experiments under controllable SNR. As shown in Figure 7, two images consisted of 30 co-circular segments as objects and 100 or 150 random segments as background. With the decrease of the SNR, difficulties in discriminating objects from background increase, which can be observed in saliency maps. When the SNR reaches 1∶5, the object is still detectable by the attention model with invariant coding, but the model without invariant coding fails.

We changed the number of segments in objects and background to obtain testing images with different SNRs. Under certain SNR, we generated 20 images with the same object but different background, and statistical results of figure-ground segregation on these synthetic images are shown in Table 1. A correct detection is obtained when among the first 10 most salient locations, those falling on the objects are not less than 60%. It can be computed by a Bernoulli binomial probability distribution that the random probability is not higher than 0.0569.

From Table 1, we can see that the correct detection rate first increases and then decreases as the number of background segments increases. When the number of background segments is small, their saliency may surpass that of objects, for the background segments are so sparse that they contrast strongly with surroundings. When the number of background segments is large, their probabilities of forming collinear or continual structures increase, leading to many local maxima in saliency maps and thus disturbing the detection of objects. In our experiments, the correct detection rate reaches maximum at SNR of 30∶100 and 40∶100.

To further analyze the contribution of invariant representation to saliency detection, we quantitate the relationship between reconstruction errors (L2 norm) and coding length (L0 norm) and plot it in Figure 8. The reconstruction error means residual error when we reconstruct a source image using the learned overcomplete topological basis set and neurons' responses.

$$\varepsilon = \| I - \phi a \|^2 ,$$

where ε is the reconstruction error, I is an image patch, ϕ is the learned overcomplete topological basis set which directly inputs to invariant representation. a is neurons' responses as computed in eq. (1). The lower ε is, the better the basis set encodes input images. The coding length (L0 norm) is the number of active neurons, i.e. the number of nonzero responses, normalized by the total number of neurons. As indicated by the sparse coding strategy in primary visual cortex, most neurons keep inactive to a stimulus while only a small fraction of neurons are activated.

As shown in Figure 8, the basis set learned by our model (which can represent invariant features) achieves higher coding efficiency than SparseNet [19] (which does not represent invariant features) and TICA [11] (a competitive model which can represent invariant features). More specifically, at the same coding length, our model can encode an

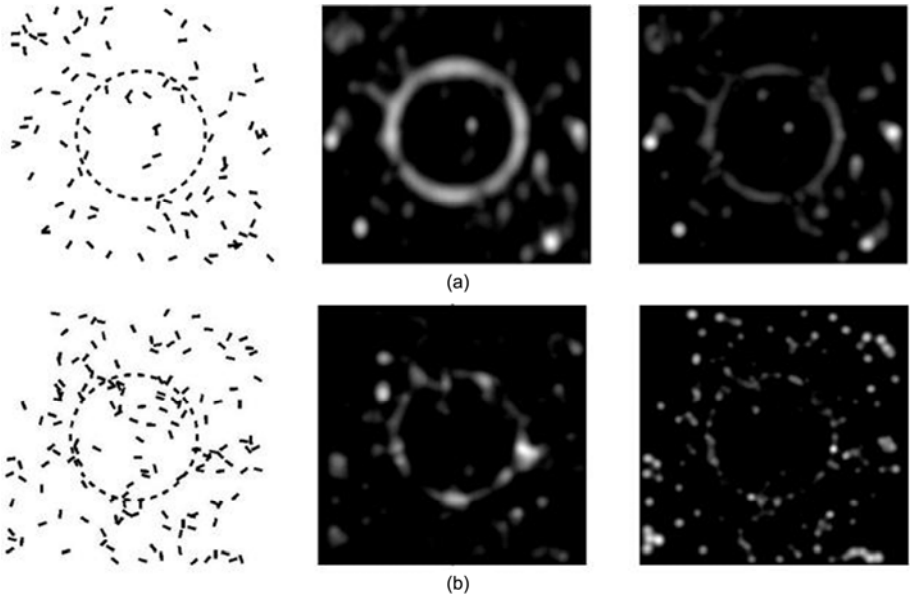

Figure 7 Saliency detection performances of the two models on figure-ground segregation under different SNRs. (a) Thirty co-circular segments in background consisting of 100 segments whose positions and orientations are random; (b) thirty co-circular segments in 150 background segments. Left: source images; mid: saliency maps of the attention model with invariant coding; right: without invariant coding. In a saliency map, the more salient part is indicated by the brighter and whiter region.

Table 1 Correct detection rates of figure-ground segregation by the attention model with invariant coding (model 1) and without invariant coding (model 2) under different SNRs

Object number: background number	Model 1	Model 2	Random
40:80	0.35	0.35	0.0569
40:100	0.90	0.85	0.0297
40:150	0.65	0.45	0.0071
40:180	0.20	0.20	0.0034
30:80	0.40	0.35	0.0242
30:100	0.80	0.70	0.0111
30:150	0.55	0.30	0.0022
30:180	0.25	0.20	0.0009

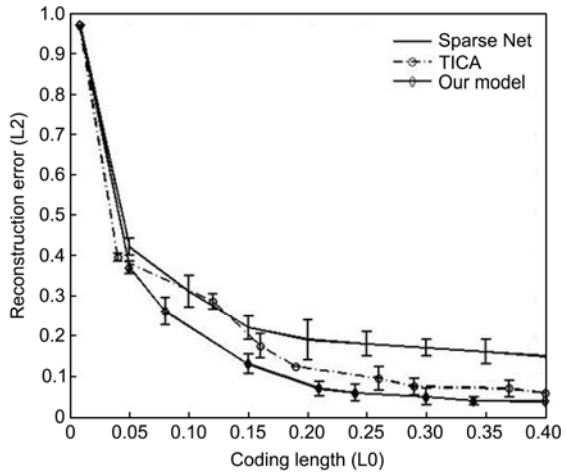

Figure 8 Coding efficiency of SparseNet, TICA and our model.

image with lower reconstruction error than the model without invariant representation and other similar models. That means our model can reserve more accurate and abundant information of inputs at the same cost, and consequently can provide better inputs for saliency detection of bottom-up attention. On the other hand, when the reconstruction error keeps constant, our model can encode information with shorter coding length. That means our model has higher coding efficiency, which is consistent with the role of bottom-up attention–to solve the information processing bottleneck. Therefore, the coding model with invariant representation is more advantageous to provide efficient input to saliency detection.

3.2 Effects of overcomplete representation on saliency detection

To investigate the effects of overcomplete representation on saliency detection, we trained different numbers of filters (bases) and produced invariant feature descriptors by pooling the filters, and then constructed bottom-up attention models based on these descriptors. The numbers of filters are set to be 100, 196, 392, and 576, respectively, and correspondingly 16, 25, 50, and 64 invariant feature descriptors by pooling the filters are selected. We compare their saliency detection performances on natural images. The testing dataset is collected by Bruce et al. [31], which includes 120 color images and eye movements from 20 observers when they view these images. The human eye tracking data can be used as a physiological basis to compare with the saliency maps obtained from attention models. Several examples are given in Figure 9.

We computed the receiver operator curve area (ROC area), a common measure in signal detection [32], to compare the performances of different models with observations from humans, and list the ROC scores in Table 2. The larger score means better consistency with human observers. Considering that the filters and invariant feature descriptors are learned from gray images and no color information is encoded, the attention models based on these can detect saliency mainly caused by intensities and orientations. Therefore, we transformed the color testing images into gray images and removed the images whose saliency was only caused by color contrasts. The remaining 82 images were used for testing.

As shown, the saliency detection accuracy improves as the number of bases increases. This shows that the more overcomplete basis set describes features of images (namely frequencies, orientations, and positions) more adequately, and provides inputs more advantageous to bottom-up attention. When the number of bases is too small, such as 100 bases pooled into 16 invariant feature descriptors, they cannot describe an image adequately, resulting in greater divergence from human detection. When the number of bases reaches 392 or 576, the change of ROC scores is tiny. This indicates that 392 bases (two times overcomplete basis set) are near saturation.

On an in-depth analysis of distributions of overcomplete bases, we find that they exhibit global topography. From the four times overcomplete bases set, we select some examples of bases in a neighborhood and show them in Figure 10. Each small square corresponds to one basis, and each group belongs to a neighborhood. The four groups represent different features of receptive fields, namely phases, frequencies, orientations, and positions. We can see features vary smoothly in local neighborhoods. On the whole, the array shows clear topological structures. Pooling such a group of receptive fields will produce invariance to corresponding feature. Besides, overcomplete bases cover the parameter space (phase, frequency, orientation, and position space) more adequately than complete bases.

Some specific non-CRFs (nonclassical receptive fields) also appear in overcomplete bases. Figure 11 lists some examples of end-inhibition, side-inhibition and curvature-selective receptive fields. Different from classical receptive fields which are coding for edges, non-CRFs are coding for corners, T-shaped stimuli or other 2D shapes. It is reported

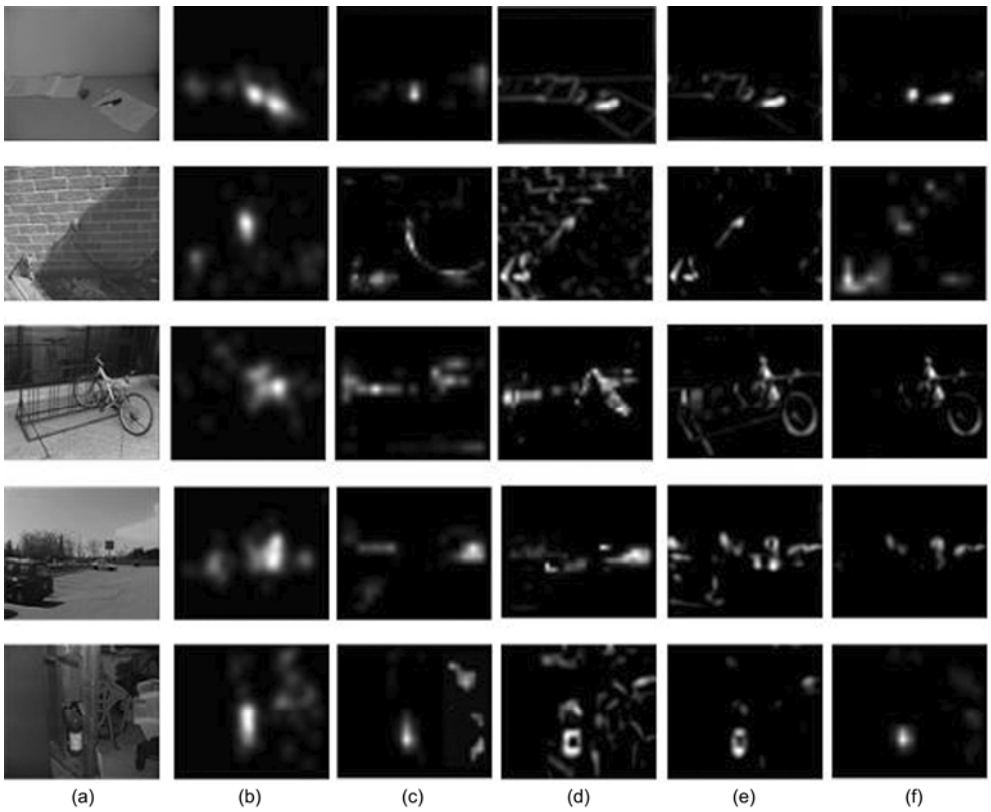

Figure 9 Saliency maps of attention models with different numbers of bases. (a) Source images; (b) human eye tracking; (c) saliency maps obtained by the attention model with 100 bases; (d) with 196 bases; (e) with 392 bases; (f) with 576 bases.

Table 2 ROC scores of attention models with different numbers of bases

Bases number	100	196	392	576
ROC	0.5722	0.6427	0.6830	0.6864

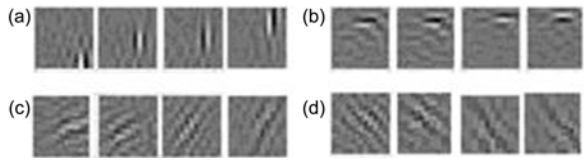

Figure 10 Examples of (a) positions, (b) phases, (c) orientations and (d) frequencies of bases in neighborhoods selected from four times overcomplete bases. Features of the bases in a neighborhood vary smoothly.

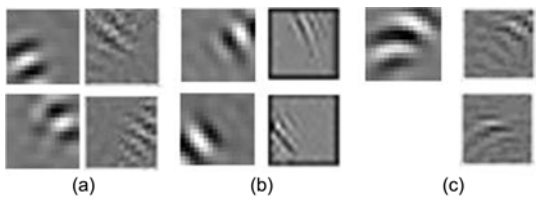

Figure 11 Non-CRFs in overcomplete bases. (a) Left: an ideal example of end-inhibition receptive fields pair used in neurophysiological experiments; right: an end-inhibition receptive fields pair learned by our model; (b) left: an ideal example of side-inhibition receptive fields pair; right: a side-inhibition receptive field pair learned by our model; (c) left: an ideal example of curvature-selective receptive field; right: two curvature-selective receptive fields learned by our model.

that end-inhibition and side-inhibition receptive fields account for 24% orientation-selective cells [33]. In complete bases sets, these non-CRFs sometimes emerge scatteredly, but mostly they do not emerge at all. In overcomplete bases sets learned by our model, the non-CRFs gather in neighborhood and achieve invariance after being pooled.

4 Discussion

In this section, we return to the two questions proposed at the beginning. First, does invariant representation in the primary visual cortex affect bottom-up attention? Second, what is the effect of overcomplete representation on saliency detection?

4.1 Invariant representation versus bottom-up attention

From multiple experiments, we can draw a preliminary conclusion-invariant representation improves saliency detection of bottom-up attention. It improves robustness to distractors, realizes illusory contour detection (named as contour completion in cognitive science) and helps to detect salient structures in spite of noises and clutters.

Chikkerur et al. [16] have analyzed the relation between object recognition and attention. It is pointed out that attention improves recognition by isolating the object of interest

from crowding and clutter. While our experimental results show invariant representation achieved in early cortical area (V1) improves the bottom-up attention effect—saliency detection.

(1) From the biological perspective, information stream from V1 transfers through V2 and V4, where the object-based attention works [34], to IT, where objects are recognized. Invariance in V1 helps attention to be relatively insensitive to transformations and preclude some noises and distracters before it sends information to upper brain area. A few theories [35,36] have suggested that such invariance in early area (V1) may underlie stability of subsequent mechanism and the whole visual perception.

(2) From the cognitive psychological view, if a group of visual primitives appear nonaccidentally (such as forming a robust structure, keep stable under small disturbance), the group is easy to pop out, attract focus of attention and waiting for further processing [28]. Invariant representation strengthens robustness of a structure, so it benefits to saliency detection.

(3) From the perspective of coding efficiency, our model can encode an image with lower reconstruction error, or at the same reconstruction error our model can encode information with shorter coding length. The reason behind this is that when we get invariant representation by pooling a group of topological bases in neighborhood, intra-group high-order dependencies are reserved while inter-group dependencies are removed. As a result, low reconstruction error and high coding efficiency guarantees high quality input for saliency detection.

4.2 Overcomplete representation versus saliency detection

The effect of overcomplete representation on saliency detection can be analyzed from two levels. On the system level, overcomplete bases encode a rich repertoire of natural image features: simple features (like localized, oriented, band-pass bars), conjunctive features, and other 2D shape features. Compared with complete bases, they represent an image more adequately and comprehensively, so they can improve saliency detection accuracy. On the single cell level, the global topography in phases, frequencies and orientations reveal advantages of overcomplete bases. The global map of phases takes on a random distribution. Pooling a group of cells with similar orientation and frequency selectivity but different (random) phase selectivity, we can get a translation invariant response. The classic energy model in V1 just get translation invariance in this way [14]. The global map of orientations shows strong correlation among bases in neighborhood. Orientation selectivity of adjacent cells keeps similar or smooth changing. This provides rotation invariance for the cell which receives input by pooling such a neighborhood. The global map of frequencies shows similarity in neighborhood, what's more, the map covers all the possible frequencies in natural images, so overcomplete bases can describe features from coarse scale to fine scale. Pooling adjacent bases can get relative invariance to scale. These invariant representations make our model robust to noises, therefore benefit to saliency detection.

5 Summary

To research which factors in coding models affect saliency detection, we construct a coding model satisfying neurobiological constraints to provide input to the bottom-up attention model. By quantitatively changing the coding constraints, we conducted experiments on images used in cognitive psychology and natural image sets to compare the effects on the saliency detection performance caused by the different coding constraints. The results of our experiments show that invariant coding and overcomplete representation are beneficial to saliency detection in bottom-up attention.

In summary, our results suggest that hierarchical invariant coding and overcomplete representation might be a general principle in visual attention and possibly in other perceptual systems.

This work was supported by the National Natural Science Foundation of China (60902058, 60975078, 61105119), Beijing Natural Science Foundation (4112047) and Fundamental Research Funds for the Central Universities (2012JBM026,2011JBZ005). We thank S. Sarkar for invaluable help and suggestions, and reviewers for commenting and suggestions.

1 Treisman A M, Gelade G. A feature-integration theory of attention. Cogn Psychol, 1980, 12: 97–136
2 Itti L, Koch C. Computational modeling of visual attention. Nat Rev Neurosci, 2001, 2: 194–203
3 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Patt Anal Mach Intell, 1998, 20: 1254–1259
4 Gustavo D, Edmund T R. A Neurodynamical cortical model of visual attention and invariant object recognition. Vision Res, 2004, 44: 621–642
5 Yan J, Zhu M, Liu H, et al. Visual Saliency detection via sparsity pursuit. IEEE Signal Proc Lett, 2010, 17: 739–742
6 Zhaoping L. Theoretical understanding of the early visual processes by data compression and data selection. Network Comp Neural, 2006, 17: 301–334
7 Maunsell J H R, Treue S. Feature-based attention in visual cortex. Trends Neurosci, 2006, 29: 317–322
8 Suder K, Worgotter F. The control of low-level information flow in the visual system. Rev Neurosci, 2000, 11: 127–146
9 Koch C, Ullman S. Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiol, 1985, 4: 219–227
10 Meur O L, Callet P L, Barba D, et al. A coherent computational approach to model the bottom-up visual attention. IEEE Trans Patt Anal Mach Intell, 2006, 28: 802–817
11 Hyvarinen A, Hoyer P. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Res, 2001, 41: 2413–2423
12 Wiskott L. How Does Our Visual System Achieve Shift and Size Invariance? Chapter 16 in 23 Problems in Systems Neuroscience. New York: Oxford University Press, 2006
13 Gao D, Mahadevan V, Vasconcelos N. On the plausibility of the discriminant center-surround hypothesis for visual saliency. J Vision,

2008, 8: 1–18
14 Grigorescu C, Petkov N, Westenberg M A. Contour detection based on nonclassical receptive field inhibition. IEEE Trans Image Process, 2003, 12: 729–739
15 Chikkerur S, Serre T, Tan C, et al. What and where: A Bayesian inference theory of attention. Vision Res, 2010, 50: 2233–2247
16 Chikkerur S, Serre T, Poggio T. Attentive processing improves object recognition. Technical Report. Cambridge, MA: Massachusetts Institute of Technology, 2009
17 Liu T, Yuan Z, Sun J, et al. Learning to detect a salient object. IEEE Trans Patt Anal Mach Intell, 2011, 33: 353–367
18 Palmer S E. Modern Theories of Gestalt Perception. Understanding Vision. New York: Blackwell, 1992
19 Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996, 381: 607–609
20 Hyvarinen A, Hoyer P. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Comput, 2000, 12: 1705–1720
21 Wang Z, Huang Y, Luo S, et al. A Biologically Inspired System for Fast Handwritten Digit Recognition. In: Benoit M, Peter S, eds. Proceedings of the 18th IEEE International Conference on Image Processing, Sept. 11-14, Brussels, Belgium, 2011
22 Elliffe M C M, Rolls E T, Stringer S M. Invariant recognition of feature combinations in the visual system. Biol Cybern, 2002, 86: 59–71
23 Nothdurft H C, Gallant J L, van Essen D C. Response modulation by texture surround in primate area V1: Correlates of "popout" under anesthesia. Vision Neurosci, 1999, 16: 15–34
24 Zenger B, Sagi D. Isolating excitatory and inhibitory nonlinear spatial interactions involved in contrast detection. Vision Res, 1996, 36: 2497–2513
25 Itti L, Koch C. A comparison of feature combination strategies for saliency-based visual attention systems. SPIE Human Vision Electron Imag IV, 1999, 3644: 373–382
26 McAdams C J, Maunsell J H R. Attention to Both Space and Feature Modulates Neuronal Responses in Macaque Area V4. J Neurophysiol, 2000, 83: 1751–1755
27 Bolz J, Gilbert C D. Generation of end-inhibition in the visual cortex via interlaminar connections. Nature, 1986, 320: 362–365
28 Palmer S E. Vision Science-Photons to Phenomenology. Cambridge, MA: MIT Press, 1999
29 Kapadia M K, Westheimer G, Gilbert C D. Spatial distribution of contextual interactions in primary visual cortex and in visual perception. J Neurophysiol, 2000, 84, 2048–2062
30 Polat U, Mizobe K, Pettet M W, et al. Collinear stimuli regulate visual responses depending on cell's contrast threshold. Nature, 1998, 391: 580–584
31 Bruce N, Tsotsos J. Saliency based on information maximization. Adv Neural Infor Process Syst, 2005, 18: 155–162
32 Tatler B, Baddele R, Gilchrist, I. Visual correlates of fixation selection: Effects of scale and time. Vision Res, 2005, 14: 643–659
33 DeAngelis G C, Freeman R D, Ohzawa I. Length and width tuning of neurons in the cat's primary visual cortex. J Neurophisiol, 1994, 71: 347–374
34 Sun Y, Fisher R. Object-based visual attention for computer vision. Artif Intell, 2003, 146: 77–123
35 Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural comput, 1992, 4: 1–58
36 Poggio T. The Computational magic of the ventral stream: Towards a theory. Nat Preced, 2011, 10: 19–59