Multifeature-Based Surround Inhibition Improves Contour Detection in Natural Images

Kai-Fu Yang, Chao-Yi Li, and Yong-Jie Li

Abstract—To effectively perform visual tasks like detecting contours, the visual system normally needs to integrate multiple visual features. Sufficient physiological studies have revealed that for a large number of neurons in the primary visual cortex (V1) of monkeys and cats, neuronal responses elicited by the stimuli placed within the classical receptive field (CRF) are substantially modulated, normally inhibited, when difference exists between the CRF and its surround, namely, non-CRF, for various local features. The exquisite sensitivity of V1 neurons to the center-surround stimulus configuration is thought to serve important perceptual functions, including contour detection. In this paper, we propose a biologically motivated model to improve the performance of perceptually salient contour detection. The main contribution is the multifeature-based center-surround framework, in which the surround inhibition weights of individual features, including orientation, luminance, and luminance contrast, are combined according to a scale-guided strategy, and the combined weights are then used to modulate the final surround inhibition of the neurons. The performance was compared with that of single-cue-based models and other existing methods (especially other biologically motivated ones). The results show that combining multiple cues can substantially improve the performance of contour detection compared with the models using single cue. In general, luminance and luminance contrast contribute much more than orientation to the specific task of contour extraction, at least in gray-scale natural images.

Index Terms—Contour detection, non-classical receptive field, surround inhibition, cue combination.

I. INTRODUCTION

CONTOUR detection is one of the fundamental tasks in computer vision applications, such as image segmentation and shaped-based object recognition [1]. Unlike edge which refers to any abrupt change in local image features such as luminance and color, contour represents changes from object to background or one surface to another [2].

A large number of methods have been proposed for edge or contour detection in recent decades. Typical examples include local differential [3], [4], phase congruency [5], [6], active contours [7], [8], statistical inference [2], [9] (more methods, see [1] for a review). Though considerable progress has been made regarding the accuracy and efficiency of contour detection, it is still challenging for the computational models to perform as intelligently as the human visual system when extracting salient contours from complex scenes [1].

Contour perception is also one of the fundamental functions of human visual system. It has been widely believed that the primary visual cortex (V1) contributes much to edge detection and contour perception [10]–[22]. The pioneering work of Hubel and Wiesel in the early 1960’s [10] reveals that the majority of neurons in V1 are exquisitely sensitive to oriented bars or edges in the classical receptive field (CRF), a limited area of visual space. Subsequently, extensive neuro-physiological findings on monkeys and cats clearly indicate a peripheral region beyond CRF, known as non-classical receptive field (non-CRF), can modulate the spiking response of a V1 neuron to the stimuli placed within the CRF [11]–[22]. The modulation of non-CRF is generally called contextual influence or center-surround interaction.

For most V1 neurons, the surround (non-CRF) exhibits inhibitory influence and the strength of inhibition is sensitive to various feature differences between the CRF and non-CRF, such as the difference in orientation [23]–[25], in spatial frequency [24], [26], in luminance, contrast and color [24], in relative moving speed [24], [27], and in spatial phase [28]. Fig. 1 shows the surround modulation properties of V1 neurons responding to different feature patterns, including orientation, luminance and luminance contrast. The neuronal responses are strongly inhibited when the stimuli within the CRF and non-CRF share similar features. The inhibition decreases or even disappears with the increment of feature difference [24], [28]. These findings suggest that the surround inhibition of V1 neurons may be an important neural basis for the visual system to perceive objects in cue-invariant ways [24].

It has been generally accepted that contextual modulation plays an important role in various visual perception tasks, such as contour integration [29]–[32], pop-out [16], [33], and figure-ground discrimination [21], [33]. Especially, contour plays an important role in advanced visual perception tasks, because it conveys much of the important information of object [1]. Although extracting contours from complex natural scenes
is difficult for the existing computer vision systems, our visual system can accomplish it quite quickly and accurately, and it has been arousing great interests among researchers of different fields to understand the underlying biological mechanisms.

Based on the findings in the physiological studies mentioned above, several biologically inspired models have been recently proposed for contour detection [34]–[44] (more methods, see [1] for a review). These models are capable of enhancing perceptually salient contours in cluttered natural scenes more effectively than some traditional edge detectors [3], [45], [46]. However, most of these biologically-inspired models are based on surround modulation and only emphasize the orientation tuning property of surround, and hence, it is difficult for them to extract the contours defined by other cues, though some of them try to improve the capability of detecting contours by introducing multi-scale strategy [36], [40], [43].

Integrating the resource of multiple scales is an efficient way to improve the contour detection [36], [40], [43], [47], [48], because the important contours are usually present at different scales with varying sizes of objects. In addition, another strategy of the visual system to cope with complex scenes is to integrate multiple visual features. In natural environments, contour detection is extremely difficult because of the complex physical conditions, such as the poor lighting, shadow, and haze. However, biological visual systems can overcome these problems and extract reliable boundaries by combining various local visual cues [49], [50]. The idea of combining multiple cues has been studied in various visual estimation problems, such as spatial feature detection [51], orientation discrimination [52], texture edge localization [53], [54] and contour detection [2], [55].

From the perspective of engineering, there are many existing algorithms that exhibit excellent capability of boundary detection by integrating the information of multiple scales or cues. Typically, Martin et al. [2] proposed the Pb method to extract and localize boundaries by taking into account multiple cues (including color, brightness and texture). Subsequently, other Pb-based algorithms have been proposed to improve the boundary detection by taking multiple scales [47], [48] or global information [48], [56]. These algorithms usually integrate the information of multiple scales or cues with a certain supervised learning method.

In this paper we are concerned with both of the contextual modulation and cue combination in the visual system, as well as the different contribution of various local features for contour detection in cluttered natural scenes. In our model, the weights of selective surround inhibition from non-CRF are first computed for each individual cue including orientation, luminance and luminance contrast. Then a scale-guided combination strategy is designed to combine the inhibition weights of different cues under the control of statistical information about CRF responses at two different scales. The combined inhibition weights are then used to modulate the strength of surround inhibition to obtain the final responses to salient edges. The source code is available at http://www.neuro.uestc.edu.cn/vccl/.

The remaining of this paper is organized as follows. In Section II, we describe the details of cue extraction, surround inhibition, weight combination of various types of inhibition, and the full multiple-cue inhibition model. In Section III, we evaluate the performance of the proposed model on RuG40 and BSDS300 datasets. Finally, we discuss our model and draw conclusions in Section IV.

II. CONTOUR DETECTION MODELS

The visual information processing begins with the photoreceptors in retina, which sample the local information of the external visual world by transforming optical signals into neural responses. In the early stages of vision, information processing is often conceived as coding the neural representation of natural scenes by a set of elementary features. Neurons in V1 can be conceived as a set of neural filters to extract specific local features in limited spatial regions [57]. Marr’s computational theories [58] suggest that local information is abstracted into the symbolic representation of objects by a series of processing stages. However, how the visual system perceives the global structures is still confusing [29], [30], [59], [60].

In this study, we propose a new model to execute the extraction and combination of local features with center-surround interaction for the specific task of contour detection.
The general framework of the proposed model is shown in Fig. 2. We first compute the responses of V1 neurons to the stimuli placed within the CRF. The three types of local cues (orientation, luminance, and luminance contrast) are extracted from the original image, and then the corresponding inhibitory weights at each location are computed based on the center-surround difference. Then, a scale-guided combination strategy (see details in Fig. 4) is designed to combine the three types of inhibitory weights into a unified one, which is then used to modulate the surround inhibition when calculating the final response at each location by subtracting the surround inhibition from the CRF response.

A. CRF Response of V1 Neuron

For the orientation-selective V1 neurons modeled in this study, we use the derivative of 2D Gaussian function to describe their properties responding to the stimuli within CRF [44], [57], [61]. The derivative of 2D Gaussian function can be expressed as

\[ RF(x, y; \theta, \sigma) = \frac{\partial g(\tilde{x}, \tilde{y}; \theta, \sigma)}{\partial \tilde{x}} \]  
\[ g(\tilde{x}, \tilde{y}; \theta, \sigma) = \frac{1}{2\pi \sigma^2} \exp\left(-\frac{\tilde{x}^2 + \gamma^2 \tilde{y}^2}{2\sigma^2}\right) \]

where \((\tilde{x}, \tilde{y}) = (x\cos(\theta) + y\sin(\theta), -x\sin(\theta) + y\cos(\theta))\), \(\theta\) is the preferred orientation of a neuron. The spatial aspect ratio \(\gamma\) and the standard deviation \(\sigma\) determine the ellipticity and the size of CRF, respectively. In this paper, we set \(\gamma\) as 0.5 on the basis of physiological finding [62].

At each location, a pool of neurons with \(N_\theta\) different preferred orientations \(\theta_i\) are employed to process the local stimuli

\[ \theta_i = \frac{(i - 1)\pi}{N_\theta}, \ i = 1, 2, \ldots, N_\theta \]  

For an input image \(I(x, y)\), the CRF response of a V1 cell with preferred orientation \(\theta_i\) is computed as

\[ e_i(x, y; \theta_i, \sigma) = |I(x, y) \ast RF(x, y; \theta_i, \sigma)| \]

where \(\ast\) denotes the convolution operation.

Then, at each location, a winner-take-all strategy is performed, i.e., the maximum CRF response over all \(N_\theta\) cells with different orientations is selected as the final CRF response, which is written as

\[ E(x, y; \sigma) = \max\{e_i(x, y; \theta_i, \sigma)\mid i = 1, 2, \ldots, N_\theta\}. \]

B. Extraction of Local Cues

In this subsection, we will describe how to extract the multiple local visual cues, i.e., local orientation \(\Theta(x, y)\), local luminance \(L(x, y)\), and local luminance contrast \(C(x, y)\).
We code the orientation information of the local stimuli around \((x, y)\) with a vector as
\[
\Theta(x, y) = [e_1, e_2, \ldots, e_{N_0}]^T_{(x, y)}
\]  
(6)

This vector will be used to compute the orientation difference of the stimuli between CRF and non-CRF.

This definition of orientation information of local stimuli at each location with a vector is quite different from that adopted in previous models [34]–[37], which normally define the preferred orientation of a neuron with the maximum CRF response over \(N_0\) neurons as the orientation of the local stimuli. Such computation of local orientation is sensitive to noise, and incorrect orientation estimate would result in ineffective surround inhibition. In contrast, we code the orientation information of local stimuli as a vector which most likely provides accurate orientation difference computation in a noise-insensitive way. This orientation coding strategy will be further demonstrated in the following subsection.

In addition, local luminance and luminance contrast are also important visual features for understanding natural scenes. Many studies have reported the statistics of local luminance and contrast in natural images, but with inconsistent conclusions. For example, Mante et al. (2005) and Frazor and Geisler (2006) claimed that local luminance and luminance contrast are independent or weakly dependent with each other in the early visual system and in natural scenes [63], [64]. However, Lindgren et al. (2008) revealed the strong spatial dependence between local luminance and luminance contrast in natural images [65]. In this work, we evaluate the contribution of local luminance and luminance contrast in contour detection with surround inhibition. Following the methods in [63]–[65], we measure the luminance and luminance contrast in local patches formed by a raised cosine weighted window
\[
w(x_i, y_i) = 0.5 \left( \cos \left( \frac{\pi}{\delta} \sqrt{(x_i - x)^2 + (y_i - y)^2} \right) + 1 \right)
\]  
(7)

where \(\delta\) is the radius of local square window \(S_{xy}\), \((x_i, y_i)\) is the location of the \(i\)th pixel in the patch with the center at \((x, y)\). In this study, \(\delta = 5\), i.e., \(S_{xy}\) has a window size of \(11 \times 11\) pixels. The local luminance and root-mean-square (RMS) defined luminance contrast are calculated according to
\[
L(x, y) = \frac{1}{\mu} \sum_{(x_i, y_i) \in S_{xy}} w(x_i, y_i) I(x_i, y_i)
\]  
(8)

\[
C(x, y) = \sqrt{\frac{1}{\mu} \sum_{(x_i, y_i) \in S_{xy}} w(x_i, y_i) \left( I(x_i, y_i) - L(x, y) \right)^2 / L(x, y)^2}
\]  
(9)

where \(\mu = \sum_{(x_i, y_i) \in S_{xy}} w(x_i, y_i)\). Note that the values of luminance and luminance contrast are linearly normalized to \([0, 1]\) (i.e., \(L, C \in [0, 1]\)) for convenience of computation.

C. Surround Inhibition of Individual Features

In this subsection, based on the maps of visual cues computed earlier, we will describe how to compute the weights of surround inhibition for individual cues at each location, i.e., the orientation-selective surround inhibition weight \(W_\Theta(x, y)\), the luminance-selective surround inhibition weight \(W_L(x, y)\), and the luminance contrast-selective surround inhibition weight \(W_C(x, y)\).

The strength of surround inhibition normally decreases with the increasing distance from the CRF center [14], [15], [18]. In this study we adopt the distance-related weighting function as [34], [35], [37], [43]
\[
W_d(x, y) = \frac{DOG_{\sigma, \rho}(x, y)}{\|DOG_{\sigma, \rho}(x, y)\|_1}
\]  
(10)

where \(\| \cdot \|_1\) denotes the \(L_1\) norm, and \(DOG_{\sigma, \rho}^+(x, y)\) is calculated as
\[
DOG_{\sigma, \rho}^+(x, y) = H \left( \frac{1}{2\pi (\rho \sigma)^2} \exp \left( -\frac{x^2 + y^2}{2(\rho \sigma)^2} \right) - \frac{1}{2\sigma^2} \exp \left( -\frac{x^2 + y^2}{2\sigma^2} \right) \right)
\]  
(11)

\[
H(z) = \begin{cases} 
0 & \text{if } z < 0 \\
1 & \text{if } z \geq 0 
\end{cases}
\]  
(12)

where \(\rho\) represents the size ratio of non-CRF to CRF. In this work we set \(\rho = 4\), which has counterpart in neurophysiology that the size of non-CRF is normally 2 to 5 times (in diameter) larger than that of CRF [18]. \(H(z)\) is used to guarantee that neuronal responses should not be negative.

On the other hand, surround inhibition strength decreases also with the increasing feature difference between the CRF and non-CRF. To effectively represent the orientation contrast between the texture patterns within CRF and non-CRF, we define the orientation difference between the center and its surround as
\[
\Delta \Theta(x, y) = \| \Theta_C(x, y) - \Theta_S(x, y) \|
\]  
(13)

where \(\Theta_C(x, y)\) is the orientation vector computed by Gaussian weighted averaging of \(\Theta(x, y)\) in the region of CRF; \(\Theta_S(x, y)\) denotes the orientation vector computed by DOG+ weighted averaging of \(\Theta(x, y)\) in the region of non-CRF.

Fig. 3 shows the examples of computing orientation difference \(\Delta \Theta\) on three types of texture patterns. \(\Delta \Theta\) is high when the stimuli within CRF and non-CRF have different orientation vectors (Fig. 3(a), high possibility to pop out), and \(\Delta \Theta\) is low when the stimuli within CRF and non-CRF have similar orientation patterns (Fig. 3(b) and (c), low possibility of being contours). In contrast, previous models, such as [37], [39], and [66], usually compute the orientation difference just based on the optimal orientation at each pixel.

Therefore, they are sensitive to noise or cluttered texture. For example, the textural pattern in Fig. 3(c) is composed of randomly oriented bars. For each bar, most of its surrounding bars show different orientations from it. Therefore, the pattern like Fig. 3(c) may be wrongly evaluated as high orientation difference.

Finally, the weight of orientation-sensitive inhibition from the surround is given by
\[
W_\Theta(x, y) = exp \left( -\frac{\Delta \Theta(x, y)^2}{2\sigma_{\Delta \Theta}^2} \right)
\]  
(14)
where the standard deviation $\sigma_{\Delta \theta}$ establishes the sensitivity of inhibitory strength with orientation difference. We experimentally set $\sigma_{\Delta \theta} = 0.2$ in this paper.

Similarly, for the features of luminance and luminance contrast, the weights of feature-selective surround inhibition from the neuron at $(x_i, y_i)$ in the non-CRF to the neuron at $(x, y)$ in the CRF are respectively given by

$$W_{\Delta l}(x, y, x_i, y_i; \sigma_{\Delta l}) = \exp\left(-\frac{\Delta L(x, y, x_i, y_i)^2}{2\sigma_{\Delta l}^2}\right)$$

$$W_{\Delta c}(x, y, x_i, y_i; \sigma_{\Delta c}) = \exp\left(-\frac{\Delta C(x, y, x_i, y_i)^2}{2\sigma_{\Delta c}^2}\right)$$

where the standard deviations $\sigma_{\Delta l}$ and $\sigma_{\Delta c}$ establish the sensitivities of inhibitory strength with the feature differences of luminance and luminance contrast, respectively. In this study we experimentally set $\sigma_{\Delta l} = \sigma_{\Delta c} = 0.05$. $\Delta L(x, y, x_i, y_i)$ and $\Delta C(x, y, x_i, y_i)$ are respectively the feature differences of luminance and luminance contrast between the stimuli at $(x_i, y_i)$ in non-CRF and $(x, y)$ in CRF, which are computed as

$$\Delta L(x, y, x_i, y_i) = |L(x, y) - L(x_i, y_i)|$$

$$\Delta C(x, y, x_i, y_i) = |C(x, y) - C(x_i, y_i)|$$

Taken together with the distance related weighting function $W_d(x, y)$ defined in (11) and the feature difference related weighting functions defined in (15), we obtain the final inhibitory weights at $(x, y)$ for the cues of luminance and luminance contrast as

$$W_L(x, y) = \sum_{(x_i, y_i)} W_{\Delta l}(x, y, x_i, y_i; \sigma_{\Delta l}) W_d(x_i - x, y_i - y)$$

$$W_C(x, y) = \sum_{(x_i, y_i)} W_{\Delta c}(x, y, x_i, y_i; \sigma_{\Delta c}) W_d(x_i - x, y_i - y)$$

(17)

where $(x_i, y_i) \in R_{nCRF}$ and $R_{nCRF}$ denotes the non-CRF region, which is defined by the DOG+ in (11).

### D. Multiple Feature Based Combination of Surround Inhibition Weights

After the surround inhibition weights of three individual features are computed by (14) and (17), we will combine them to obtain the final surround inhibition weights of the neurons responding to the input scenes. The details are as follows (Fig. 4). We compute the gradient magnitudes (i.e., CRF responses) at two different scales (a fine scale with $\sigma$ and a coarse scale with $2\sigma$) with (1)~(5), denoted by $E(x, y; \sigma)$ and $E(x, y; 2\sigma)$, respectively. Then, the difference of gradient magnitude maps between the fine and coarse scales is used to guide the combination of surround inhibition weights of three individual features according to (Fig. 4)

$$W_{\text{com}}(x, y) = \begin{cases} 
\max(W_\Theta, W_L, W_C)(x, y), & \Delta E(x, y) > 0 \\
\min(W_\Theta, W_L, W_C)(x, y), & \Delta E(x, y) \leq 0 
\end{cases}$$

(18)

$$\Delta E(x, y) = \mathbb{N}(E(x, y; \sigma)) - \mathbb{N}(E(x, y; 2\sigma))$$

(19)

where $\mathbb{N}()$ is a linear normalization operator. For noise smoothing, gaussian filtering is applied to the $\Delta E(x, y)$ map. Note that, since $\Delta E(x, y)$ is just the difference of two “blurred” responses, it is sufficient to remove some isolated points on $\Delta E(x, y)$ map with a simple Gaussian filter, though a good image denoiser, e.g., some edge-preserving image smoothing methods [67]–[70], provides valuable preprocessing for contour detection.

Why such weight combination strategy described by (18) and (19) is valid, the reason is explained as follows. Generally, the gradient magnitude map at a large scale (i.e., the CRF responses $E(x, y; 2\sigma)$) includes reliable contours, but misses the detailed edges and may be inaccurate in localization. In contrast, the CRF responses at a fine scale (i.e., $E(x, y; \sigma)$) covers more details, also including the undesired edges in cluttered or textured regions. In the $\Delta E(x, y)$ map reflecting the difference of (normalized) edge responses at two scales, the pixels with $\Delta E(x, y) > 0$ more likely belong to undesired non-meaningful edges than perceptually salient contours; hence, the weights of surround inhibition should be stronger at the locations with $\Delta E(x, y) > 0$, which is realized by selecting the maximum of $W_\Theta$, $W_L$, and $W_C$, as done in (18). In contrast, the pixels with $\Delta E(x, y) \leq 0$ more likely belong to perceptually salient contours than unwanted
detailed edges, and therefore, the weights of surround inhibition should be weaker at such locations, which is realized by choosing the minimum of $W_\theta$, $W_L$, and $W_C$, as done in (18).

Taken together, the strategy described above may endow our model with the ability to extract more perceptually salient contours and suppress more undesired trivial edge fragments, compared to the model using only single cue, considering the high complexity of natural scenes.

Note that, the idea of our “scale based” guiding described in Fig. 4 seems, to some extent, similar to that of well-known “unsharp masking” since we also utilize information at two different scales. Our strategy, however, is distinguished from others in the following ways. We utilize the difference between the two “images” at two different scales (i.e., $E(x, y; \sigma)$ and $E(x, y; 2\sigma)$), here the “images” $E(x, y; \sigma)$ and $E(x, y; 2\sigma)$ are not the images blurred by different Gaussian filters, but the neuronal responses to edges at two different scales. That is, $\Delta E(x, y)$ denotes the difference of edge responses at two scales. In contrast, the difference between the original image (one scale) and its Gaussian blurred version (another scale) in the technique of “unsharp masking” reflects the edges that need to be enhanced. In short, our $\Delta E(x, y)$ is quite different from “unsharp masking” in its meaning and role. Another point to be mentioned is that our model can not be classified into true multiresolution-based ones, since we do not directly extract edges at multiple scales, but just use the multiscale (two-scale) information to determine the surround inhibition weights before extracting final edges at one (fine) scale, as described in the following section.

E. Salient Contour Extraction

Considering the fact that the fine-scale edge information includes more details with more accurate localization, we first compute the non-specific (i.e., isotropic) surround inhibition by convoluting $E(x, y; \sigma)$, the CRF response map at the fine scale, with $W_d(x, y)$, the distance-based weighting function

$$Inh(x, y) = E(x, y; \sigma) \ast W_d(x, y)$$

Then, we construct the final neuronal responses to the salient contours by subtracting the combined weight (i.e., $W_{com}(x, y)$) modulated surround inhibition from the CRF responses at the fine scale according to

$$r(x, y) = H(E(x, y; \sigma) - \alpha \cdot W_{com}(x, y) \cdot Inh(x, y))$$

where $H(\cdot)$ is defined in (12). The factor $\alpha$ denotes the connection strength between the neurons within the CRF and its surrounding non-CRF.

III. EXPERIMENTAL RESULTS

In this section, we first give experimental results on a natural image dataset (RuG40) [34] to demonstrate how multifeature-based surround inhibition in our model offers advantages over single-feature based surround inhibition. We tested our model using the whole RuG40 dataset with a population of 40 natural images, and quantitative performance was compared with existing models based on the ground truth contour maps. We also tested our model’s performance on the BSDS300 dataset [71].

Table I summarizes the meanings of the parameters involved in our model. The parameter settings are also listed in Table I, which are identical for both RuG40 and BSDS300 datasets except $p$. In this study, we select the values of $\alpha$ as $\alpha \in [1.0, 5.5]$ (for RuG40 and BSDS300 datasets) and $p$ as $p \in [0.1, 1.0]$ (for RuG40 dataset only) to test the robustness of various models, because these two parameters exert relatively high influences on the overall performance.

A. Inhibitory Effects

When computing gradient magnitude map, we can find that many unwanted texture edges exist in the CRF response
TABLE I
THE SUMMARY OF PARAMETER MEANINGS AND SETTINGS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Equation</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>Spatial aspect ratio of the CRF filter (derivative of Gaussian)</td>
<td>(2)</td>
<td>0.5</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>The size of CRF, the standard deviation of the derivative of Gaussian filter</td>
<td>(2)</td>
<td>2.0</td>
</tr>
<tr>
<td>$N_d$</td>
<td>Number of orientations of the CRF filters</td>
<td>(3)</td>
<td>8</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Radius of local square window for extracting local luminance and contrast feature</td>
<td>(7)</td>
<td>5 (pixels)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Ratio of the two standard deviations in DOG+ model</td>
<td>(11)</td>
<td>4.0</td>
</tr>
<tr>
<td>$\sigma_{\Delta \phi}$</td>
<td>Inhibition sensitivity of the feature difference of orientation</td>
<td>(14)</td>
<td>0.2</td>
</tr>
<tr>
<td>$\sigma_{\Delta \lambda}$</td>
<td>Inhibition sensitivity of the feature difference of luminance</td>
<td>(15)</td>
<td>0.05</td>
</tr>
<tr>
<td>$\sigma_{\Delta c}$</td>
<td>Inhibition sensitivity of the feature difference of luminance contrast</td>
<td>(15)</td>
<td>0.05</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Surround inhibition factor, or the texture attenuation factor</td>
<td>(21)</td>
<td>[1.0, 5.5]</td>
</tr>
<tr>
<td>$p$</td>
<td>Fraction of candidate edge pixels that should be retained in the contour edge map during hysteresis thresholding</td>
<td>–</td>
<td>[0.1, 1.0]</td>
</tr>
</tbody>
</table>

Fig. 5. Demonstration of inhibitory effects. (a) input image, (b) gradient magnitude map at the fine scale, i.e., the neuronal responses to the stimuli within CRF, (c) the isotropic surround inhibition (with distance-related weighting), (d) the inhibitory weight combined based on multiple cues, (e) the surround inhibition modulated by the combined weight, (f) the final responses to salient contours. For (b) and (f), higher grey levels correspond to stronger neuronal responses. For (c)∼(e), higher grey levels correspond to higher inhibitory weights or surround inhibitions.

map (Fig. 5(b)). In order to suppress textures, the isotropic inhibition model (ISO) [34] computes the inhibition term by considering the distance-related influence (Fig. 5(c)). With the absence of feature-selective modulation, the isotropic inhibition suppresses texture efficiently, but meanwhile it also inhibits some perceptually salient contours, especially the contours embedded in cluttered background. In our model, the final inhibition term (Fig. 5(e)) is modulated simultaneously by the distance-(Fig. 5(c)) and cue-based weights (Fig. 5(d)). From Fig. 5(d), we can clearly see that at the locations of object contours (e.g., the contour of rhinoceros), the inhibition strengths are generally weaker; in contrast, the texture regions (e.g., the grasses) receive relatively stronger inhibition. Consequently, our inhibition strategy is more efficient for texture suppression and contour protection. Fig. 5(e) shows that the locations of contours receive weaker inhibition than other locations. Hence, the proposed model can respond strongly to perceptually salient contours, but is relatively insensitive to textures (Fig. 5(f)).

We also evaluated the performance of our models with individual-cue-based inhibition. We obtained the response of the models with individual cue inhibition by replacing $W_{com}(x, y)$ in (21) with the individual cue based inhibitory weight in (14) or (17). Note that, we use the symbol $MCI$ to denote the operator with multiple-cue inhibition, and $OI$, $LI$, and $CI$ to denote the operators with orientation, luminance, and luminance contrast inhibition, respectively. We also use $Noinh$ to denote the operator without surround inhibition, i.e., the gradient magnitude operator at the fine scale.

B. Experiments on RuG40 Dataset

We tested our model on an available contour detection dataset, RuG40 dataset [34], which includes 40 natural images and associated ground truth binary contour maps drawn by hand. The dataset can be downloaded from http://www.cs.rug.nl/~imaging/databases/contour_database/. This dataset has been widely used to evaluate the performance of contour detectors [34], [36], [38], [43], [57]. In order to quantitatively evaluate the performance of our method, the binary contour maps were constructed using the standard procedure of non-maxima suppression followed by hysteresis thresholding [3], [34]. We measured the similarity between the binary contour map detected by a model and the known ground truth (human-marked contours), according to [34]

$$P = \frac{\text{card}(E)}{\text{card}(E) + \text{card}(E_{FP}) + \text{card}(E_{FN})}$$

where $\text{card}(S)$ denotes the number of elements of the set $S$. $E$ refers to the set of contour pixels that are correctly detected by an algorithm. $E_{FP}$ and $E_{FN}$ refer to the set of false positive pixels and false negative pixels, respectively. Therefore, a higher measure $P$ reflects a better overall performance of contour detection. We compared our model with another surround inhibition-based model (isotropic inhibition, ISO) [34], and the operator without surround inhibition ($Noinh$).

In this experiment, 100 parameter combinations (10 values of $\alpha$ within [1.0, 5.5] and 10 values of $p$ within [0.1, 1.0]) were
Fig. 6. The mean performance of multiple models over all the images of the whole RuG40 dataset with various parameter combinations. Each image provides a performance measure of $P$, and each bar in the figure represents the mean $P$ over all the images with one parameter combination $(\alpha, p)$. The notion of $P(\alpha, p)$ below each panel illustrates the best mean $P$ obtained using the parameter values of $\alpha$ and $p$ on all the images.

Fig. 7. Results of various models on two example images from RuG40 dataset [34] using the optimal parameter setting derived on the whole dataset (see Fig. 6). The number on the top of each image indexes its performance ($P$ value).

Two example images are listed in Fig. 7. In general, the operator of $\text{Noinh}$ responds to abundant of edges in texture regions. Surround inhibition can suppress the textures efficiently, while $\text{ISO}$ still has weak capability of removing texture edges. In contrast, operators with cue-dependent surround inhibition ($\text{OI}$, $\text{LI}$, and $\text{CI}$) show different properties when detecting contours. For instance, $\text{CI}$ misses some low contrast defined contours, but suppresses more textures. Finally, $\text{MCI}$ usually obtains the best contour detection results by combining multiple cues. The results suggest that the integration of multiple cues resulted in improvement of performance and robustness of contour detection system.

Fig. 8 illustrates more comparisons of our $\text{MCI}$ operator with other recent biologically-inspired models, including Butterfly-shaped inhibition [72] and Adaptive inhibition [43]. For Butterfly-shaped inhibition, we selected the parameter
as $\sigma = 2.0$, and other parameters were the same as those listed in Table I in [72]. For *Adaptive inhibition*, we also set the parameters as same as ref. [43]. For these two existing methods, we took 10 different values of $p$ within $[0.1, 1]$, as done in our model. Fig. 8 clearly shows that the results with image-specific optimal parameters (red bars) are substantially better than that with the optimal parameter setting for the whole dataset (blue bars). More importantly, in both situations, the model of MCI shows obvious improvement over other models, except the models of butterfly-shaped inhibition [72] and adaptive inhibition [43] with optimal parameters for the whole dataset. The good performance of the models of butterfly-shaped inhibition and adaptive inhibition could be attributed to the dividing of the whole surrounding non-CRF into four subregions, which provides a more flexible way to adapt to local features. In addition, Fig. 9 shows more results on two example images, which indicate again that the proposed *MCI* model exhibits excellent capability of suppressing textures and detecting perceptually salient contours.

### C. Experiments on BSDS300 Dataset

We further tested our models on Berkley Segmentation Dataset (BSDS300) [71], which includes 300 images and associated 5~10 human-marked segmentations for each image. Fig. 10 presents several examples of the probability of contours computed by our model (in (21)) followed by the operation of non-maxima suppression. We used the same parameter settings as that on RuG40 dataset except $p$, because hysteresis thresholding was not used in this experiment. The first and second columns of Fig. 10 show three images from BSDS300 dataset and their human-marked boundaries. The third to fifth columns show the results of three state-of-the-art algorithms: *Global Probability of Boundary* ($gPb$) [48], $Pb(BG+TG)$ [2], and $Pb(BG)$ [2]. The last column shows the results of our *MCI* model, indicating that our *MCI* responds to less texture edges. Note that, some edges are considered as boundaries in ground truth of BSDS300. For example, the test image in the bottom row of Fig. 10 clearly shows that some subjects have marked the edges of grasses as contours, and the ground truth was constructed by overlapping all the markers of different subjects.

Quantitative comparison was also carried out using the test set of 100 images in BSDS300. We computed the so-called F-Score, defined as $F = 2PR/(P + R)$ [2]. $P$ and $R$ represent precision and recall, respectively. F-Score signals the similarity between the contours identified by the algorithm and human subjects. From Table II, our *MCI* operator gets the score of 0.62, which is higher than other surround-inhibition-based methods, including the methods proposed in this paper (*OI, LI, and CI*) and *ISO* [34]. Our *MCI* operator ($F = 0.62$) also scores higher than another recent biologically-inspired model, *PC/BC-V1+lateral+texture* [57] ($F = 0.61$). The model of *PC/BC-V1+lateral+texture* scores 0.61 under the condition that recurrent lateral excitatory connections and other special lateral connections are incorporated into the classical *PC/BC* model of V1 [57]. In contrast, our model only uses the surround inhibition, which could be regarded to be equivalent to the special lateral connections used in the model of *PC/BC-V1+lateral+texture*. In other words, our model does not introduce the similar mechanism as recurrent lateral excitation. This reveals that our model could be extended by integrating facilitatory surround modulation, considering the special role of collinear facilitation of V1 neurons in visual processing [20], [59], [73], [74] and its successful applications in computational modeling [37], [41], [42], [75], [76].

The scores of our methods are below those of some learning-based algorithms, such as $gPb$ [48], $mPb$ [47], $Pb(BG+TG)$ [2] and *BEL* [77]. However, these methods need extra supervised learning processing, and they normally employ more cues or image scales. In fact, our proposed models use only image intensity as the source of contour detection. In our methods, the multiple cues and scales are used to guide the surround inhibition to remove textures...
IV. DISCUSSION AND CONCLUSION

At the level of individual neurons, classical receptive fields (CRFs) respond to all the main edges including both desired and undesired ones, and the surround inhibition from non-CRF contributes to the suppression of undesired edges. To this point, appropriately weighting the surround inhibition is crucial. Focusing on this central aspect, this work proposed a new framework for contour detection based on the surround inhibition in multiple visual feature dimensions, with emphasis on how to determine the surround inhibition weights with the resource of multiple features.

The main contribution of the work could be summarized as follows. (a) A center-surround interaction based framework was designed to integrate various features for the task of contour detection. To our knowledge, this is the first attempt to build a unified model combining multiple features with the
center-surround mechanism of the visual system. Other visual features could be easily integrated into this unified framework in the future. (b) A new strategy was designed to weight the surround inhibition by combining the surround inhibition of individual features. This strategy employs the difference of neuronal responses at two different scales, which provides an intuitive and effective way to determine where most probably exist the salient contours and undesired edges. (c) Our results show that anisotropic (feature-selective) surround inhibition works better than isotropic surround inhibition when extracting salient contours. Note that the work in [34] shows opposite conclusion. We attribute the better performance of anisotropic model to our new orientation coding scheme which, to some extent, could code texture patterns. (d) Our results also show that compared to the widely used feature of orientation, luminance and luminance contrast provide much more contributions to contour detection, at least on the grey-level images. This suggests that low level cues play different roles in contour perception in the visual system. The luminance and luminance contrast perhaps play a major role in suppressing textures and detecting existing contours. Meanwhile, higher-order statistics (e.g., orientation) may convey additional information that potentially augments the first-order statistics.

In this work, we only employed the visual features of orientation, luminance, and luminance contrast. Though these three features have been widely used in the field of image processing, the contribution of the work does not lie in the introduction and extraction of these three specific features, but how different visual features contribute cooperatively to the visual perception task of contour detection in a unified way of center-surround interaction. In addition, the implementation of computing surround inhibition weights using (14)∼(17) faithfully simulates the physiological observations (no detailed analysis was given here, but the involved functions like EXP and DoG have been widely used in the fields of neurovision). As for the computational integration of various surround inhibition weights, it seems somewhat heuristic in (18)∼(19), since its biological substrates is quite unclear. Even so, the implementation is still quite promising, considering the fact that the operations of Min and Max involved in (18)∼(19) have been widely accepted as the canonical neural computations in the visual system [78].

Besides the three features employed, the differences of other features, such as spatial frequency and color [24] and spatial phase [28], between the CRF and non-CRF also selectively activate V1 cells. This finding indicates another extension to our current framework in the future by combining more visual features, which is an efficient way that has been widely adopted [2], [50]–[56]. The main reason that color was not considered in this work is that most color information is processed with special color-opponent mechanisms (e.g., the various types of double-opponent neurons in V1). Our recent work [44] studied the possible function of a kind of double-opponent color neurons in detecting boundaries. To effectively utilizing the cue of color, the current framework needs to be extended by adding another branch for non-center-surround mechanism based processing. Another important question is how to integrate the signals carried by various information sources. Compared to others [2], [47], [48], our model is distinguished mainly by how to use and how to integrate the multiple cues, i.e., our model uses multiple cues to generate possible surround inhibition weights, and these cue-specific surround inhibition weights are combined using a Min and Max based strategy; the combined weight is used to modulate the final neuronal response by weighting the surround inhibition. To our knowledge, such strategy of fusing multiple cues is the first attempt, along the lines of both biologically- and non-biologically-inspired. For example, some algorithms integrate the multiple cues/scales by a linear sum operator with a supervised learning [2], [47], [48], and generally, a preprocessing phase before learning based cue combination is required to manually tune those cue-specific parameters in order to obtain the best performance. In contrast, our model only needs a preprocessing phase to handle a list of parameters (Table I). In general, different images may require different optimal parameter values (see Fig. 8). However, with one set of parameters statistically optimal at the level of entire dataset (RuG40), we have still obtained competitive performance in comparison to other models. Our model acted relatively robust, i.e., parameter values in broad range results in similar acceptable performance (see Fig. 6). This allow us to select proper parameter values effortlessly.

In summary, in this study we proposed a center-surround interaction based model by combining multiple local cues to improve contour detection in cluttered scenes. The contribution of the work to the field of image processing is founded on the fact that more and more research in contour detection has been focusing on understanding and modeling of the biological mechanisms that human use to perceive contours [1]. It is expected that the work is not only of interest to neurovision researchers who care of unclear biological substrates, but also quite relevant to the field of image processing, where models of higher efficiency and efficacy are especially desired.

ACKNOWLEDGMENT

The authors would like to thank Prof. De-Wen Hu, Hong-Mei Yan and Dr. Ling Wang for valuable conversations and discussions on the presented work. They would like to thank Dr. Ting Li and Kaiwen Cheng for comments on manuscript. Thoughtful comments and suggestions from anonymous reviewers for improving the manuscript are also very much appreciated. They would also like to thank Nicolai Petkov and his colleagues for their source codes used for validating our models.

REFERENCES
