Optik 126 (2015) 522–529

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A biologically inspired computational approach to model top-down and bottom-up visual attention

Longsheng Wei[a,b,*], Dapeng Luo[a]

[a] Faculty of Mechanical and Electronic Information, China University of Geosciences, Wuhan 430074, China,
[b] School of Automation, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 22 January 2014
Accepted 16 January 2015

Keywords:
Visual attention
Biologically inspired
Top-down
Bottom-up
Saliency map

ABSTRACT

A biologically inspired computational approach to model top-down and bottom-up visual attention is proposed in this paper. This model includes a training phase and an attention phase. In the training phase, low-level visual object's feature dimensions such as color, intensity, orientation and texture are used; the visual features are extracted from object itself and do not depend on the background information. These features are represented by mean and standard deviation stored in long-term memory (LTM). In the attention phase, corresponding features are extracted in the attended image. For each feature, the similarity map is obtained by comparing training feature map and attended feature map. The more similarly, the stronger of the similarity map. Then all the similarity maps are combined into a top-down saliency map. In the same time, a bottom-up saliency map is acquired by the contrast of attended image itself. At last, top-down and bottom-up saliency map are fused into a final saliency map. Experimental results indicate that: when the attended object does not always appear in the background similar to that in the training images or their combinations change hugely between training images and attended images, our proposed approach is excellent to VOCUS top-down approach and Navalpakkam's approach.

1. Introduction

The visual system requires attention and guidance of the attention because the eyes provide the central nervous system with more information than they can process [1]. Attention has been classified into two types based on whether its deployment over a scene is primarily guided by scene features or volition: one is often called bottom-up and is mainly driven by low-level processes depending on the intrinsic features of the visual stimuli; the other refers to knowledge-based top-down processes [2]. The top-down attention is more complex to model because it needs to represent object in LTM [3–5] and uses the memory to detect likely target object in attended scenes [6–8].

Most computational models [9–16] of visual attention are bottom-up and are inspired by the concept of feature integration theory [17]. The most popular is the one proposed by Itti et al. [18] and it has become a standard model of bottom-up visual attention, in which salience according to primitive features such as color, intensity and orientation are computed independently. There are also many top-down visual attention models [19–26], which can attend the target object quickly using prior knowledge. Both the attention types just emphasize a part of attention, so several models [27–30] integrate both top-down and bottom-up attention. Well-known computational models include Visual Object detection with a CompUtational attention System (VOCUS) [31] and the model proposed by Navalpakkam and Itti [32]. VOCUS top-down approach integrates both top-down and bottom-up attention. The top-down part is based on a weight of target object and its background for each feature in training image. The weight is used to combine different feature conspicuity map in test image. Conspicuity maps represent the level of saliency for single visual feature. Navalpakkam et al. also combine bottom-up attention and top-down attention. The top-down component uses accumulated statistical knowledge of the visual features of the desired search target object and its background, to optimally tune the bottom-up maps so that target detection speed is maximized. The performances of these two approaches are influenced by scenes, so they are not efficacious when the background of object changes hugely.

In order to address the above problem, we propose a biologically inspired computational approach to model top-down and bottom-up visual attention in this paper. We just use some low-level feature dimensions such as color, intensity, orientation and texture. In the training phase, all the visual feature types are extracted from object itself and they are irrelevant to background information. These

* Corresponding author. Tel.: +86 13437184026.
E-mail address: weilongsheng@163.com (L. Wei).

http://dx.doi.org/10.1016/j.ijleo.2015.01.004
0030-4026/

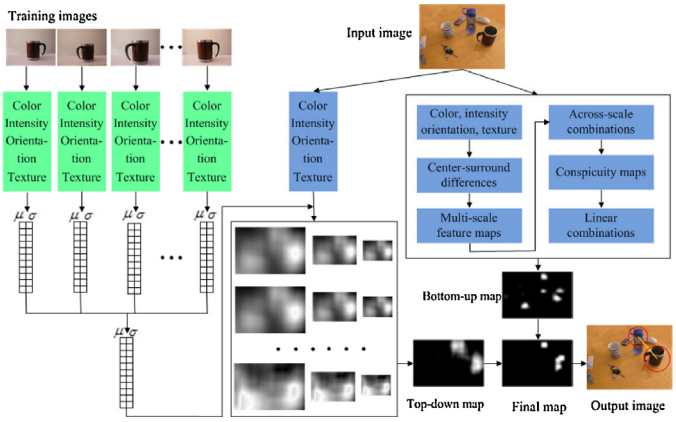

Fig. 1. Our proposed approach: given a task such as "find a red cup in the input image". Firstly, target's feature types are extracted from training images and are represented by mean and standard deviation. Secondly, this information is used to compare the similarity in the input image and form a top-down saliency map. Thirdly, bottom-up saliency map is acquired by the contrast of attended image itself. At last, top-down and bottom-up saliency map are fused into a final saliency map, which are guided attention to likely target locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

feature types are represented by mean and standard deviation. In the attention phase, corresponding feature types are extracted in the attended image. For each feature, the similarity map is obtained by comparing training feature map and attended feature map. The more similarly, the stronger of the similarity map. Then all the similarity maps are combined into a top-down saliency map. In the same time, a bottom-up saliency map is acquired by the contrast of attended image itself. Then, top-down and bottom-up saliency maps are fused into a final saliency map. At last, the size of each salient region is obtained by maximizing entropy. Experimental results indicate that: when the attended object does not always appear in the background similar to that in the training images or their combinations change hugely between training images and attended images, our proposed approach is excellent to the VOCUS top-down approach and Navalpakkam's approaches. Our proposed approach is shown in Fig. 1.

The remainder of this paper is organized as follows. Section 2 presents the object representation including feature extract and the training of object representation. While attentional selection is described in Section 3, this part introduces how to acquire saliency maps and how to acquire the size of salient region. Section 4 shows experimental results, and Section 5 concludes this paper.

2. Object representation

In this part, basic image feature dimensions such as intensity, color, orientation and texture are considered and all the visual feature types are extracted from object itself and do not depend on background. The target object occurs is looked as random variable which follows a normal distribution. Therefore, every feature type is represented by mean and standard deviation.

2.1. Feature extraction

Retinal input is processed in parallel by low-level feature dimensions including color, intensity, orientation and texture. Each feature dimension is divided into some different feature types such as color is divided into red, green and blue three feature types denoted as r, g and b. For each feature type, we represent the object by calculating the mean and standard deviation in this feature type.

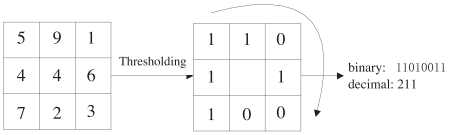

Fig. 2. The LBP operator.

For example, $(\mu_{i,1},\sigma_{i,1})$, $(\mu_{i,2},\sigma_{i,2})$ and $(\mu_{i,3},\sigma_{i,3})$ represent red, green and blue feature types of the ith training object, respectively.

We divided intensity into intensity on (light-on-dark) and intensity off (dark-on-light). The reason is that the ganglion cells in the visual receptive fields of the human visual system are divided into two types: on-center cells respond excitatory to light at the center and inhibitory to light at the surround, whereas off-center cells respond inhibitory to light at the center and excitatory to light at the surround [33]. In this paper, we convert the color input image into gray-scale image to obtain an intensity image $I=(r+g+b)/3$ and let center/surround contrast be intensity on, that is to say: let the difference of each pixel value and its surround average pixel value as the corresponding value (negative values are set to zero). In same way, surround/center contrast be intensity off, $(\mu_{i,4},\sigma_{i,4})$ and $(\mu_{i,5},\sigma_{i,5})$ represent intensity on and intensity off as be described above.

There are four orientations in our model: 0°, 45°, 90° and 135° [34,35]. The orientations are computed by Gabor filters detecting bar-like features according to a specified orientation. Gabor filters, which are the product of a symmetric Gaussian with an oriented sinusoid, simulate the receptive field structure of orientation-selective neurons in primary visual cortex [33]. A Gabor filter centered at the 2-D frequency coordinates (U,V) has the general form of

$$h(x,y) = g(x',y')\exp(2\pi i(Ux'+Vy')), \tag{1}$$

where

$$(x',y') = (x\cos(\phi)+y\sin(\phi), -xsin(\phi)+y\cos(\phi)), \tag{2}$$

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x^2}{2{\sigma_x}^2}-\frac{y^2}{2{\sigma_y}^2}\right). \tag{3}$$

σ_x and σ_y are the scale parameters, and the major axis of the Gaussian is oriented at angle ϕ relative to the x axis and to the modulating sinewave gratings. In this paper, let the scale of Gabor filters equal to the scale of training object and let ϕ equal to 0°, 45°, 90°and 135°, respectively. We represent this four feature types by calculating the mean and variance as $(\mu_{i,6},\sigma_{i,6})$, $(\mu_{i,7},\sigma_{i,7})$, $(\mu_{i,8},\sigma_{i,8})$ and $(\mu_{i,9},\sigma_{i,9})$.

For texture feature, we consider local binary pattern (LBP) [36], which describes the local spatial structure of an image and has been widely used in explaining human perception of textures. Ojala et al. [37] first introduced this operator and showed its high discriminative power for texture classification. At a given pixel position (x_c,y_c), LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its eight surrounding pixels (Fig. 2). The decimal form of the resulting 8-bit word (LBP code) can be expressed as follows:

$$\mathrm{LBP}(x_c,y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n. \tag{4}$$

where i_c corresponds to the gray value of the center pixel (x_c,y_c), i_n to the gray values of the 8 surrounding pixels, and function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{5}$$