Optik 125 (2014) 2313–2319

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Novel biological visual attention mechanism via Gaussian harmony search

Junnan Li, Haibin Duan*

State Key Laboratory of Virtual Reality Technology and Systems, School of Automation Science and Electrical Engineering, Beihang University (BUAA), Beijing 100191, PR China

ARTICLE INFO

Article history:
Received 19 May 2013
Accepted 14 October 2013

Keywords:
Harmony search algorithm
Visual attention
Saliency map
Pre-training

ABSTRACT

Most of the visual attention models are based on the concept of a two-dimensional saliency map, which encodes the conspicuity of the object in the visual scene. The visual attention model proposed by Laurent Itti is used in this work. In Itti's model, the saliency map is calculated via combining the information across several modalities, including color, intensity, and orientation. In this work, we propose a pre-training process to select the weightings used in the combining of feature maps to make the target more conspicuity in the saliency map. Harmony search (HS) algorithm is used in the pre-training process to obtain the weightings. HS is a new heuristic algorithm, which mimics the improvisation of music players. Its performance has been verified by many benchmark problems. We modify the pitch adjustment process of the original HS to improve the optimization performance and accelerate the convergence rate. The modified algorithm is named Gaussian harmony search (GHS).

1. Introduction

When biologic visual systems try to interpret a visual scene, it is impossible to process all the visual stimuli occurred at the retina at one time, and then a mechanism called visual attention is employed. In this mechanism, some circumscribed regions in the scene will be spotlighted for further analysis [1]. It has been proved by anatomy that most of the early visual processing areas in brain participate in the development of visual attention. Visual attention mechanism is crucial for organism because it enables them to detect targets quickly in the visual environment [2].

A large amount of research has been conducted by previous researchers to interpret this mechanism. The basis of many attention models dates back to Treisman and Gelade's Feature Integration Theory. Koch and Ullman then proposed a feed-forward model to combine these features to obtain a saliency map. In around 1980's to 1990's, scientists proposed a classical framework for visual attention. It suggests that the visual attention mechanism is driven by two basic cues [3,4]. One is a bottom-up cue, it is based on the image itself, no characteristics of the target are previously known and used in the attention development. The other one is a top-down cue, it is based on the characteristics of the targets in the visual scene, and objects with certain characteristics will be paid attention to. Some visual stimuli are intrinsically conspicuous in a given scene. For example, a bright red coat is on a bed with white sheet, and then the coat will attract people's attention. In this case, saliency is primarily driven by a bottom-up cue. The red coat is intrinsically conspicuous in the background. There is another example: a person is looking for a white cup on the shelves. Then his or her attention will focus on those white colored, cylinder shaped things. In this case, saliency is primarily driven by a top-down cue. The white cup is not intrinsically conspicuous in the background. But it stands out because some characteristics of it have previously been known by the people who are looking for it.

In 1998, Itti proposed a computer implementation for the bottom-up scheme visual attention based on Feature Integration Theory [5]. Other methods like discriminant hypothesis [6], spectral models [7], sparse and efficient coding [8], Bayesian and graphical models [9] were also proposed by researchers based on information theory. Itti's bottom-up visual attention model is well known and most frequently used by other researchers. In Itti's model, a saliency map is obtained via combining information across several modalities, including color, intensity, and orientation. When this model is applied to specific target detecting tasks, it is crucial to select the weightings used in the combining process to make targets more conspicuity in the saliency map. However, in a bottom-up scheme, no information about the targets is previously known, so the detecting tasks may fail if the targets are not intrinsically conspicuous in the background. Recently, more researchers are interested in top-down visual attention. Top-down models are divided into three major types: visual search models, context models, and task-driven

* Corresponding author at: Beihang University, State Key Laboratory of Virtual Reality Technology and Systems, Beijing 100191, PR China. Tel.: +86 10 8231 7318.
E-mail address: hbduan@buaa.edu.cn (H. Duan).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.10.075

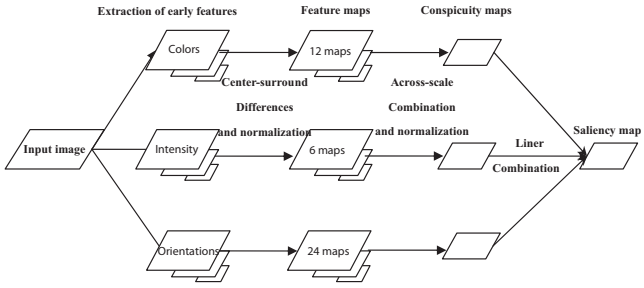

Fig. 1. Architecture of Itti's bottom-up model.

models [10]. However, the majority of studies on top-down attention are at the analysis level, and there is not a widely accepted computational model.

We propose a pre-training process, which is proposed to obtain the weightings used in the combination of feature maps. A series of training images containing the target in complicated backgrounds are used in the pre-training process. We create a cost function to evaluate the conspicuity of the target in the saliency maps of the training images; the weightings used in the combination are seen as inputs of the cost function. In this way, the problem of selecting the weightings is translated into a mathematical problem. HS is applied to solve this optimization problem. In recent years, many intelligence algorithms have been successfully applied to solve optimizing problems in the field of computer visual [11–14].

HS is a relative new heuristic algorithm raised by Z. W. Geem and J. H. Kim in 2001. They are inspired by the improvisation process of music players. Musicians first improvise a new music harmony with their memory and adjust some of the syllables, then, evaluate the new music harmony from an aesthetic point of view. This process can be imitated to create a heuristic algorithm. The performance of HS has been verified by many benchmark problems. In many cases, HS has several advantages over other popular optimization algorithms. For instance, in Genetic Algorithm (GA), to generate a new input vector, only two parents in the existing input vectors are considered, while HS consider all existing input vectors with a possibility. Furthermore, HS could also generate a totally new vector independent to the existing vectors. These features endow HS a better ability to escape from local optimums and find a better solution.

In this paper, we combine the 'bottom-up' cue with the 'top-down' cue by adding a pre-training process into the previous bottom-up scheme. This process obtains a set of weightings used in the combination of feature maps. Targets are supposed to be more conspicuity in the saliency map using the new scheme. The rest part of this paper is organized as follows. In Section 2, we will introduce some basic information about Itti's bottom-up visual attention model. In Section 3, the previous HS is first introduced, and then we propose a new method named GHS. In Section 4, we give a detail description of the pre-training process. We apply our new scheme to several cases to test its performance, the comparison results and analysis can be seen in Section5.

2. A bottom-up visual attention model

Visual information processing system is a limited system. Visual information that reaches the retina is a lot more than the system could process. Visual attention mechanism helps organism to focus their attention on certain areas in the visual scene. It has been a great challenge for computational neuroscience scientists to develop a computational model to interpret how visual attention is developed. In 1998, Itti and his partners proposed a bottom-up model for visual attention (Fig. 1). The model is concise and works well in many cases. Therefore, it is widely accepted by other researchers. Fundamental of this model is introduced in this section. More introduction and applications of the bottom-up model can be seen in [15–21].

2.1. Extraction of pre-attention features

Computation of a series of pre-attention features of the entire image is the first step in this model. The pre-attention features include color, intensity, and orientations. The input of this step is a static two-dimensional color image, each pixel in the input image contains such information as (x, y, r, g, b), where (x, y) is the location of the pixel in the image, r, g, b are the values of the red, green, and blue channel respectively [15,16].

A Gaussian pyramid of nine spatial scales is created by subsampling and applying low-pass filters to the input image. Therefore, the pyramid contains nine images of different scales $\sigma \in \{0, 1...8\}$, where 0 represents the original image, and 8 represents the smallest scale. In the next step, image of each feature is computed by a set of liner center-surround operation, which is an operation between fine center scales and coarse surround scales. Center scales are defined as $c \in \{2, 3, 4\}$, surround scales are defined as $s = c + \delta$, $\delta \in \{3, 4\}$.

An intensity image I can be obtained by $I = 1/3(r + g + b)$. A Gaussian pyramid $I(\sigma)$ is created based on I. Then, a set of feature maps $I(c,s)$ are obtained based on $I(\sigma)$. $I(c, s)$ contain six maps; it is obtained by Formula (1). $I(c,s)$ is related with the color contrast of the input image.

$$I(c, s) = \left|I(c) - I(s)\right| \tag{1}$$

where, $c \in \{2, 3, 4\}$, $s = c + \delta$, $\delta \in \{3, 4\}$. The subtraction operator in Formula (1) represents that we first resize $I(s)$ to the scale of the corresponding image $I(c)$, and then operate point-by-point subtraction in the two images.

Human's visual attention mechanism deals with color information based on four-color channels: red, blue, green, and yellow channel. As a consequence, four color channels are defined, $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green, $B = b - (r + g)/2$ for blue, and $Y = (r + g)/2 - |r - g|/2 - b$ for yellow. Then four Gaussian pyramids $R(\sigma)$, $B(\sigma)$, $G(\sigma)$, $Y(\sigma)$ are obtained based on the four channels.

Another two sets of feature maps $RG(c,s)$ and $BY(c,s)$ are obtained by Formula (2). They are related with the color contrast of the input image

$$\begin{cases} RG(c, s) = \left|(R(c) - G(c)) - (C(s) - R(s))\right| \\ BY(c, s) = \left|(B(c) - Y(c)) - (Y(s) - B(s))\right| \end{cases} \tag{2}$$

The last set of feature maps is based on the orientation information. A set of Gabor filters are applied to the intensity image I, creating the orientation pyramid $O(\sigma, \theta)$, where, $\sigma\theta$ $\{0, 1...8\}$ and $\theta \in \{0°, 45°, 90°, 135°\}$. The orientation feature maps encode the orientation contrast between the center scales and surround scales. $O(c, s, \theta)$ is computed according to Formula (3).

$$O(c, s, \theta) = \left|O(c, \theta) - O(s, \theta)\right| \tag{3}$$

In total, 42 feature maps are obtained in this step, including 6 intensity feature maps, 12 color feature maps, and 24 orientation feature maps.

2.2. Combination of feature maps

In the previous step, 42 feature maps are obtained. The following question is that how to process the 42 feature maps to create a unique saliency map [22]. First, feature maps in each channel