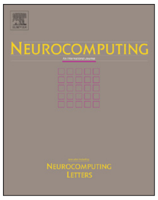
Neurocomputing 133 (2014) 1–17

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Mimicking visual searching with integrated top down cues and low-level features

Jiawei Xu*, Shigang Yue

School of Computer Science, University of Lincoln, Brayford Pool, Lincoln LN6 7TS, United Kingdom

ARTICLE INFO

Article history:
Received 15 October 2013
Received in revised form
16 November 2013
Accepted 30 November 2013
Communicated by Hantao Liu
Available online 28 January 2014

Keywords:
Visual searching
Learning bases
Top-down cues
Motion sensitive neurons
Multi-layer neural network

ABSTRACT

Visual searching is a perception task involved with visual attention, attention shift and active scan of the visual environment for a particular object or feature. The key idea of our paper is to mimic the human visual searching under the static and dynamic scenes. To build up an artificial vision system that performs the visual searching could be helpful to medical and psychological application development to human machine interaction. Recent state-of-the-art researches focus on the bottom-up and top-down saliency maps. Saliency maps indicate that the saliency likelihood of each pixel, however, understanding the visual searching process can help an artificial vision system exam details in a way similar to human and they will be good for future robots or machine vision systems which is a deeper digest than the saliency map. This paper proposed a computational model trying to mimic human visual searching process and we emphasis the motion cues on the visual processing and searching. Our model analysis the attention shifts by fusing the top-down bias and bottom-up cues. This model also takes account the motion factor into the visual searching processing. The proposed model involves five modules: the pre-learning process; top-down biasing; bottom-up mechanism; multi-layer neural network and attention shifts. Experiment evaluation results via benchmark databases and real-time video showed the model demonstrated high robustness and real-time ability under complex dynamic scenes.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The role of visual searching in everyday tasks is ubiquitous and important to human beings and primates. The visual searching process can be traced back to two levels, the visual attention and visual searching process. The former one is to identify the most important and informative part of a scene which saves the processing costs and reduces the burden of computation. It also can be thought as the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things. Visual searching is a type of perceptual task requiring attention that typically involves an active scan of the visual environment for a particular object or the target among other objects or the distractors. One of the essential components on visual searching is the attentive guidance of attention towards a target item. In a cluttered visual scene, humans are able to find what they are looking for with amazing efficiency. Therefore, understanding where people to look can help an artificial vision system to exam details in a way similar to human as they will be good to future robots or machine vision systems which are a deeper digest than the saliency maps.

The study on visual searching can be traced back 2 decades ago. Wolfe [1] proposes the guided search model via the neuropsychology experiments. In the visual searching process, attention will be directed to the item with the highest priority. If that item is rejected, then attention will move on to the next item and the next, and so forth. The guided search theory follows that of parallel search processing. Another popular way of methods is the construction on the computational models, such as Itti's bottom-up attention model on mimicking the visual attention and attention shifts [2], or the integrated models like Kanan [3]. These works study on the features cues extraction and integration by using natural statistics or mathematical modeling. Further details will be described in Section 2.1.

The state-of-the-art research focuses on the modeling of the visual cortex and represents their function by using neurophysiological model and other similar approaches, such as experiments on the monkey and mus musculus, this model explains the visual activities effectively using bottom-up and top-down hypothesis. Comparing with our model, Itti's model is a bottom-up approach without considering the top down cues thoroughly. Secondly, it is not well integrated with the motion vectors under dynamic scenes, as these cues affect a lot on our visual searching processing and attention. Thirdly, even Itti's [2] model did simulate the attention shifts and attentional regions in their model, our approach puts the emphasis on the priority on the current frames or images against with the complex background, which is stressed

* Corresponding author.
E-mail addresses: jxu@lincoln.ac.uk (J. Xu), syue@lincoln.ac.uk (S. Yue).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.037

by more circles while the upcoming attended regions are represented by fewer circles shown by our model, which leads an intuitive and straightforward comprehension to readers. On the other hand, Kanan's model [3] integrated the top-down and bottom-up cues together however it is used for saliency computing, while our model mimics the searching processing to study the visual mechanism submerged in their perception.

This paper attempts to build a hierarchical perception system based on the top-down cues and bottom-up mechanism. A number of neurophysiologic experiments [4–11,19,51–54] has been performed on biased competition neural mechanisms which is consistent with the theory of Duncan and Humphreys [12] (i.e., with a role for a top-down memory target template in visual search). The biased competition hypothesis proposes that multiple stimuli in the visual field activate populations of neurons that engage in competitive interactions. Biased competition serves to prioritize task relevant information to make visual searching more efficient. A large amount of visual information is taken in at any given moment and there is a limited capacity available for processing. The visual system therefore needs a way to select relevant information and ignore irrelevant stimuli. Each cell in neural network represents the attended stimulus, thereby suppressing cells representing distracting stimuli. Based on these characteristics, our proposed model involves five modules: the pre-learning process, top-down biasing, bottom-up mechanism, multi-layer neural network and attention shifts. From the pre-learning level, we categorize our dictionary into the target and non-target templates. The top-down stage extracts the features from the templates and yields the bias in the weight tuning process. The bottom-up visual attention mechanism is the selective tuning processing which is driven by the properties of the proto-objects. The multi-layer neural network integrates the weight tuning and generates the human eye fixations and attended regions via its iterative learning procedures. Each layer reflects its response map in visual cortex. The attention shift follows the routine of winning-take-all (WTA) and inhibition of return (IOR). The extensive comparisions with other latest visual searching models have verified the effectiveness of our proposed model.

The rest of the paper is organized as follow: in Section 2 we propose an efficient visual search framework. The top-down and bottom-up cues, multi-layer neural network and attention shift are described in Section 3. The experimental results and performance evaluation are reported in Section 4. Then we discuss some interesting phenomenon and compare other models in Section 5. Finally, Section 6 concludes the paper and gives the outlook.

2. The proposed model

According to the recent studies, the visual search models can be categorized into two attributes. The first attribute is based on the parallel search analysis; the second attribute is based on the serial search analysis [13]. The following sections elaborate on the visual search model and describe our proposed attention model briefly. The details of the proposed model will be further illustrated in Section 3.

2.1. *Related work on visual searching*

Tracing back to the last century, the first computational attention model from Itti [2], and the visual saliency model has been developed in a fast and multiple ways. A lot of researches have been applied into many ways, such as image segmentation, object recognition, image retrieval, etc.

Visual searching is a type of perceptual task requiring attention which typically involves an active scan of the visual environment for a particular object or feature (the target) among other objects or features (distractors). Visual search can take place either with or without eye movements. The ability to consciously locate an object (target) among a complex array of stimuli (distractors) and has been extensively studied over the past 40 years. Practical examples of this can be seen in everyday life such as picking out a book from the bookshelf, animals searching for food among piles of leaves, and finding your friend in a large crowd of people. Many visual search paradigms have used eye movements as a means to measure the degree of attention given to stimuli [13,14]. However, vast research [15,16] to date suggests that eye movements move independently of attention and therefore it is not a reliable method to examine the role of attention. Many behavioral researches focusing on reaction time have dominated this field of research. Many of the previous literatures on visual searching use reaction time (RT) in order to measure the time taken to detect the target among its distractors. In our experiments, we also compare other state-of-the–art methods, such as Kanan, Yang, Bruce, Xie [5,17,18,20] and our model shows a better performance with these methods.

2.2. *Proposed model*

As shown in Fig. 1, our proposed model is a hierarchical and multi-model architecture trying to mimic the visual searching process when objects come into the vision field. First, we start training the target relevant template (positive) and irrelevant template (negative) beforehand. This is a data acquisition and normalization stage. Second, after features deduction of the target, the model yields the top-down biasing with specific pre-given task. Another pathway is the bottom-up mechanism, when the agents receive a visual input; it will generate low-level features and could be integrated with the top-down biasing. After bias competition and weight tuning, the multi-layer neural network serves as neural perception layer in visual cortex. The output is focus on attention and moves to next location according to the mechanism IOR [21] and WTA [22]. This process is called attention shift in terminology.

Fig. 2, however, defines our model in a two directional ways. The pre-learning stage is the top-down cues origin, which we set the interesting objects as the agents' targets. Reciprocally, the agents process it as the top-down bias (displayed in red flows). Another pathway is induced by visual input, after pre-attentive processing, such as feature extraction and normalization, the bottom-up mechanism (shown in green flows) affects the attention selection. The final end is the attention shift which means a new round of visual search.

In detail, we assume that the top-down cues generates the long-time memory which affects the later processing, while the bottom-up mechanism yields the short-time memory, it merges into the feature integration into the neural network model. In order to simplify the problem, we temporarily disregard the short-time memory during the attention selection stage. The information in long-time memory can keep for a couple of weeks or more long time, which includes motion skill and details, etc. We map long time memory into the attention selection stage, which is reflected by the pre-learning cues, such as intensity, texture and edge. This will be described in Section 3.

Figs. 1 and 2 represent same meanings. The only difference is to explain our model from the detailed steps and directional procedure. In the following sections, we will elaborate on our visual searching model.

3. The visual searching model

The first stage of our model constructs a large proto-object based dictionary, such as human face, cat, car and snorting valve.

Download English Version:

https://daneshyari.com/en/article/6866531

Download Persian Version:

https://daneshyari.com/article/6866531

Daneshyari.com