Signal Processing 147 (2018) 92–100

Contents lists available at ScienceDirect

Signal Processing

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

Predicting human gaze with multi-level information

Xiaoning Zhang, Di Xiao, Jianhua Li, Jinqing Qi, Huchuan Lu*

School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116023, China

ARTICLE INFO

Article history:
Received 29 July 2017
Revised 7 December 2017
Accepted 2 January 2018
Available online 10 January 2018

Keywords:
Attention
Eye fixation
Feature
Adaptive weight learning

ABSTRACT

Eye fixation models, which try to quantitatively predict human eye attended areas in visual fields, have received increasing interest in recent years. In this paper, a novel framework is proposed for the detection of eye fixations. First, a multi-channel detection module, which extracts information of color contrast, salient object proposals and center bias from input image, is conducted to introduce various useful information into the subsequent fixations detection. In salient object detection channel, we employ the multi-instance learning (MIL) algorithm to determine which object proposal can attract attention, which avoids the fuzzyness of positive sample selection. Second, an adaptive weighted fusion method achieved by deep learning framework is proposed to fuse the multi-level information (i.e., contrast, objective, center bias) together for the detection of fixations, so that the integration of information between each level becomes more scientific. Finally, the detection result is optimized by embedding semantic information. Experimental results show that the algorithm has achieved competitive results in MIT1003, MIT300 and Toronto120 dataset.

1. Introduction

When observing an image, human's eyes are attracted by visual information of interest in the image according to biological evidence [1]. Because of the differences in the subjective perceived quality of different regions in the scene image, certain regions stand out from their adjacent background areas and attract attention. Saliency detection, which simulates human attention mechanism, tries to find regions of interest in a variety of images. Eye fixation detection is an important branch of saliency detection and often applied as pre-operation in image segmentation [2,3], adaptive image compression [4], image retrieval [5,6], objects tracking [7], object recognition [8], etc. Processing the image according to visual saliency can reduce a large extent of computational complexity without losing important information.

Two hypotheses were proposed in the development of models for eye fixations prediction. Most computational models follow the saliency map hypothesis [9] which defines saliency as outliers of the distribution of visual features in the image. The regions with rare features are assigned high saliency values and attract human's attention. Consistent with biological findings [1], this kind of models mostly operate on low-level features, such as orientation, color and intensity. Another hypothesis [10] which is object-based claims that objects may better help predict fixations and directly attract attention. It is difficult to determine which of these two hypothesis is more accurate, or we can say they are both correct considering the difference of contents in the images. Through the research and analysis of a large number of natural scene images, we simply summarize human's attention mechanism as follows: First, when observing an image, human's eyes will be attracted by visual information of interest such as person, face, text, billboard and so on (see first row in Fig. 1). Second, if there is no aforementioned visual information, the observer tends to observe the salient regions which usually have a noticeable contrast to the rest of the image (see second and third rows in Fig. 1). Last but not the least, if the image is homogeneous or very complex without information and areas mentioned above, the observer tends to look at the center of the image (see fourth row in Fig. 1). We present some examples to illustrate human's attention mechanism in Fig. 1.

Based on the research and analysis above, we design our eye fixations detection model. A multi-channel detection framework, where different channels correspond to different levels of information, is adopted in our work. In low level information channel, color information is exploited to detect the regions which have strong contrast to the rest of the image. In salient object detection channel, objects which can attract human's attention are detected through sample classification. However, when using a linear SVM to classify the samples, there exists positive sample fuzziness problem. To address this problem, multi-instance learning (MIL) algorithm is applied into our model, which leads to better detection results. In addition, based the content of the image, observers tend

* Corresponding author.
E-mail address: lhchuan@dlut.edu.cn (H. Lu).

https://doi.org/10.1016/j.sigpro.2018.01.003
0165-1684/© 2018 Elsevier B.V. All rights reserved.

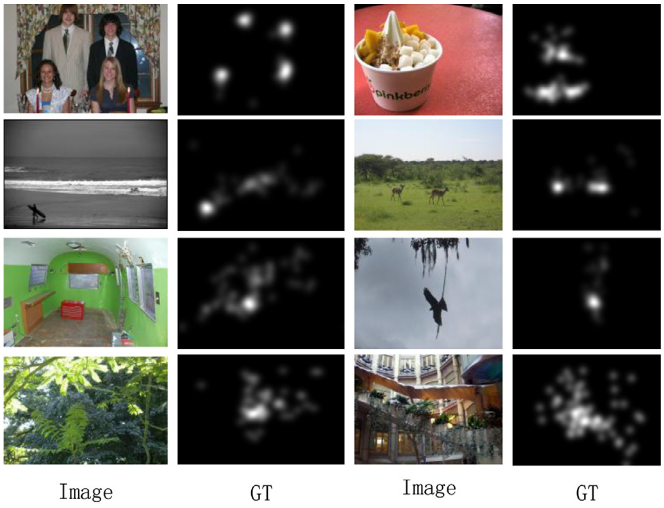

Fig. 1. Some examples to illustrate human's attention mechanism.

to look at the center of the image more or less. To take this kind of inclination into consideration, we add another channel for center bias. To make the integration of the information from each channel more reasonable, we propose an adaptive weighted fusion method achieved by deep learning framework to judge the reliability of the maps from each channels. We use the output of the network to determine the fusion weights of each channels. At last we add three kinds of semantic detections to further optimize the final results. Extensive experimental results demonstrate the proposed method has achieved competitive results in MIT1003 and other databases.

In summary, we have three main contributions:

1. The multi-instance learning (MIL) algorithm is applied in eye fixations detection to determine whether the target in the image is salient, avoiding the fuzzyness of positive sample selection.
2. We propose an adaptive weighted fusion method achieved by deep learning framework to fuse multi-level information, so that the integration of information between each level becomes more scientific.
3. The proposed algorithm achieves state-of-the-art results on the MIT1003 and other datasets.

The rest of this paper is organized as follows. Section 2 introduces the related work of our algorithm. Section 3 describes the details of our proposed model. Section 4 reports the experimental results and discussions of our model. Finally, we draw conclusions in Section 5.

2. Related work

The current fixation algorithm can be divided into two categories: bottom-up generative model and top-down discriminant model. As for generative model, some computational visual attention models follow the Feature Integration Theory [9,11]. Itti model [12] proposed by Koch and Ullman constructs Gaussian pyramid for three feature channel: brightness, color and direction respectively. Each feature is computed by the center-surrounded operator to generate three saliency maps which are added linearly to obtain the final saliency map. Hou et al. [13] propose a model using Fourier transform spectral residuals for saliency detection. This model uses a low-pass filter to filter the amplitude spectrum to obtain redundant information and then removes it from the original image to obtain saliency map. In addition, AWS [14,15] and RARE [16] have also achieved good detection effects. The former uses decorrelation and distinctiveness to compute saliency while the later is mainly based on regional color rarity.

Different from bottom-up methods, most top-down models are achieved under the framework of supervised learning. Liang et al. [17] propose an eye fixation detection model which utilizes higher-level information. They combine sift feature with the BOW model to calculate color and shape saliency maps. Then they use object detectors named Object Bank to obtain multi-object detection maps and train SVM to get the weights of object detection maps as the high-level eye fixation map. Xu et al. [18] propose a new eye fixation detection architecture, integrating the information at three levels: pixel level, object level and semantic level. Experimental results show that semantic information played a very important role in eye fixation detection. However, the semantic information is calibrated manually in their experiments. In recent years, some deep learning methods using Convolutional Neural Network (CNN) have also achieved good results, such as Mr-CNN [19] and e-DN [20]. Moreover, in some object detection works such as [21] and [22], different semantic information and CNN-based network are all exploited for better performance. To take advantage of multi-level information and deep learning networks, we propose a model to incorporate multi-channel detection module, CNN-based adaptive weighted fusion and semantic information embedding.

Another related work we need to mention is multi-instance learning algorithm which plays an important role in the multi-channel detection module by solving the problem of inaccurate sample selection. The concept of multi-instance learning was proposed by Dietterich in the study of drug activity prediction [23] in the 1990s. Since the concept of multi-sample learning has been put forward, this learning method has been regarded as the most potential machine learning method. The training set for multi-instance learning consists of bags containing a number of instances, each of which has a training label. If at least one of the instances in a bag is positive, the bag is marked as positive, whereas the bag is marked as negative. The goal of multi-instance learning is to learn a classifier through training data and classify the subsequent bags using the classifier correctly. Babenko et al. [24] improve the multi-instance learning object detection algorithm and use it in tracking. They think during the updating procedure of the traditional classifier, problems of sample fuzziness widely exist and this kind of bias will become more and more serious during the iteration process. In order to solve this problem, multi-instance classification strategy is exploited. Image blocks near the tracking results can be regarded as instances and form a bag together. Using this bag to update the classifiers can avoid sample fuzziness. In [25], a new multi-instance learning algorithm RMI-SVM is proposed. They treat the positiveness of instance as a continuous variable, use Noisy-OR model to enforce the MIL constraints, and jointly optimize the bag label and instance label in a unified framework. Moreover, they apply RMI-SVM to a common object discovery and achieved good results. Therefore, in our work, we make use of RMI-SVM that avoids sample fuzziness problem to determine whether a detected object proposal is salient or not.

3. Proposed model

In this section, we elaborate our proposed approach. The architecture of proposed model is showed in Fig. 2. First, we propose a multi-channel detection framework to encode three levels of information (i.e., salient objective, color contrast and center bias) for eye fixation detection. Then a CNN-based fusion network is presented to incorporate multiple channel information. Finally, we add three semantic detection to improve the performance of the final results.