Pattern Recognition Letters 49 (2014) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Markov chain based computational visual attention model that learns from eye tracking data ☆

Ma Zhong [a,*], Zhao Xinbo [b], Zou Xiao-chun [c], James Z. Wang [d], Wang Wenhu [a]

[a] Laboratory of Contemporary Design and Integrated Manufacturing Technology, Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[c] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
[d] School of Computing, Clemson University, Clemson, SC 29634, USA

ARTICLE INFO

Article history:
Received 11 February 2013
Available online 18 June 2014

Keywords:
Attention model
Eye tracking
Markov chain
Saliency
Human fixations

ABSTRACT

Computational visual attention models are a topic of increasing importance in computer understanding of images. Most existing attention models are based on bottom-up computation that often does not match actual human attention. To address this problem, we propose a novel visual attention model that is learned from actual eye tracking data. We use a Markov chain to model the relationship between the image feature and the saliency, then train a support vector regression (SVR) from true eye tracking data to predict the transition probabilities of the Markov chain. Finally, a saliency map predicting user's attention is obtained from the stationary distribution of this chain. Our experimental evaluations on several benchmark datasets demonstrate that the results of the proposed approach are comparable with or outperform the state-of-art models on prediction of human eye fixations and interest region detection.

1. Introduction

The Human Visual System (HVS) can quickly select the most important and informative portions of a scene for further analysis. This ability enables us to allocate limited processing resources on the most relevant visual input. Understanding which part of an image will attract human attention has benefited a wide range of applications such as advertising design, seam carving [23], as well as compression [29] and object recognition [24] and more. Therefore, modeling the visual attention to predict where human will look has long been of interest in the computer vision community.

The process of obtaining a computational visual attention model can be described as follows: Given an image, we want to measure the likelihood of a location to attract the attention of human observers without any background knowledge. Most existing visual attention models use a set of biologically plausible linear filters, e.g., Gabor or center-surround (DoG) filters as a front-end, the outputs of which are non-linearly combined into a real number that indicates visual saliency [10]. Although these models have the ability to measure the salience of a location, they are of limited use. This is because only the image features are taken into consideration, therefore the generated salience maps do not always match actual human fixations, and the performance of the models largely depends on tuning many parameters [31].

Since gaze fixations reflect the true visual attention of a human, a promising way to improve the visual attention model is to exploit the actual eye tracking data. In this paper, we have proposed a novel visual attention model that is learned from actual eye tracking data. First, we extract low-level features and high level features of the image. Next, we use a Markov chain to model the relationship between the image features from different part of an image and the saliency. The transition probabilities of the chain are learned from actual eye tracking data using support vector regression (SVR). Finally, the saliency of a query image can be estimated from the stationary distribution of the trained Markov chain. The whole process is shown in Fig. 1.

The rest of the paper is organized as follows. The related work on computational visual attention models is reviewed in Section 2. Section 3 introduces the experimental setting for collecting true eye gaze data. The extraction of low level and high-level features of the image is introduced in Section 4. In Section 5, we propose a new computational visual attention model based on true eye gaze data by using a Markov chain. Experimental results are reported in Section 6. Finally, concluding remarks are drawn in Section 7.

☆ This paper has been recommended for acceptance by J.K. Kämäräinen.
[*] Corresponding author. Tel.: +86 13571814259.
E-mail address: mazhong@mail.com (M. Zhong).

http://dx.doi.org/10.1016/j.patrec.2014.06.002
0167-8655/© 2014 Elsevier B.V. All rights reserved.

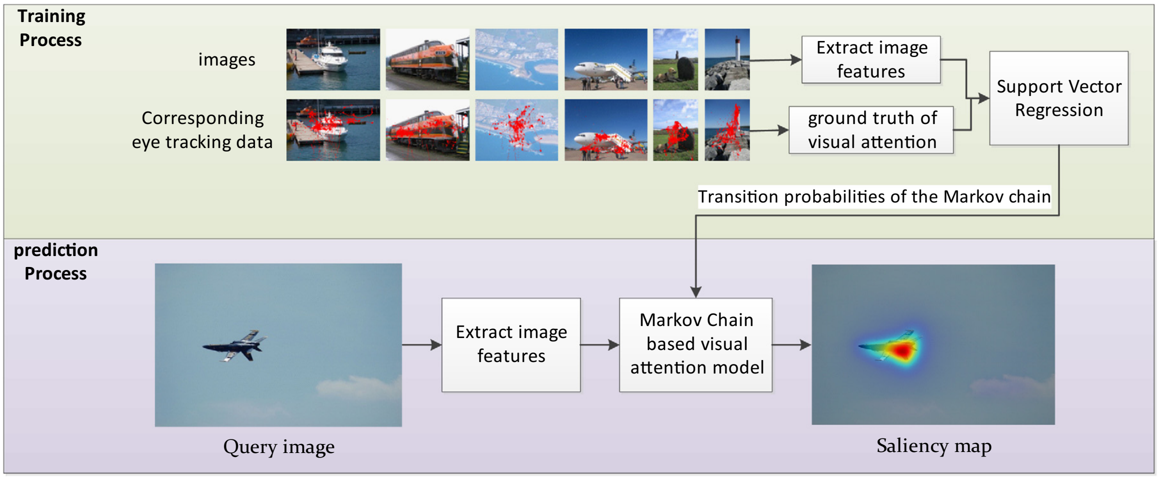
Training Process
images
Corresponding eye tracking data
Extract image features
ground truth of visual attention
Support Vector Regression
Transition probabilities of the Markov chain
prediction Process
Extract image features
Markov Chain based visual attention model
Query image
Saliency map

Fig. 1. The diagram of the training and prediction process of the computational attention model.

2. Related work

A number of computational models of visual attention have been developed and the state of the art is rapidly improving. The majority of existing models can be divided into the following categories.

2.1. Bottom-up models

The computational model of visual attention was first introduced by Itti et al. [15]. Specifically, they proposed using a set of feature maps from three complementary channels as intensity, color, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Based on this, many other researchers suggested improvements. Draper and Lionelle [5] introduced selective attention as a front end (SAFE) which modified the original approach such that it is more stable with respect to geometric transformations like translations, rotations, and reflections. Walther and Koch [28] extended the Itti model to attend to proto-object regions. Meur et al. [20] adapted the Itti model to include the features of contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. Privitera and Stark [22] have improved the Itti model by adding features of symmetry. The above approaches are based on local image features, such as center-surround operations, that have trouble with activating salient regions distant from the object borders, even if one varies over many choices of scale differences and combinations thereof. Harel et al. [14] exploited the computational power, topographical structure, and parallel nature of the Markov chain model to achieve natural and efficient saliency computations of their graph based visual saliency model. The model estimates the saliency in a global way, such that it can robustly highlight salient regions, even far away from the object borders.

2.2. Top-down models

Besides bottom-up factors, top-down factors also play an important role in HVS. In order to obtain a better simulation of the human's attention, the bottom-up and the top-down saliency have to be fused to obtain a single focus of attention [10]. Some researchers have made some headway on this. Gao et al. [12] proposed a unified framework for top-down and bottom-up component as a classification problem with the objective being the minimization of classification error. They first applied this framework to object detection in which a set of features are selected such that a class of interest is best discriminated from all other classes, and saliency is defined as the weighted sum of features that are salient for that class. Torralba et al. [26] proposed a Bayesian framework for the task of visual search. It combines low-level salience and scene context to guide the search. Another way to add top-down component to models is to use object detectors. Cerf et al. [3] indicated that faces and text strongly attract attention. They add a conspicuity map indicating the location of faces and text to the Itti model, and show that it improves the ability to predict eye fixations in natural images.

2.3. Models that utilize eye tracking data

Since eye tracking data of humans provide the ground truth for human visual attention, some researchers try to make use of them. Kienzle et al. [17] learn a visual saliency model directly from human eye movement data using a support vector machine (SVM). Judd et al. [16] did a similar job, but they also used high-level features. Both methods show better performance than the traditional methods that purely depend upon image features. This demonstrates the ability of using actual eye tracking data to improve the visual attention model. Nevertheless, they directly use local feature as attributes to train a SVM, missing the connection between local and peripheral features. Liang et al. [18] refined a region based attention model [11] with actual eye tracking data using a Genetic Algorithm (GA), their results show that the refined model outperforms the original one, for which only the image features are used. However, their attention model is still based on low-level image features, and only the parameters are optimized by eye tracking data, which give only limited improvement in performance.

3. Eye tracking data acquisition

The eye gaze data we used for training data is from Judd et al. [16], referred to as the MIT data set, for our saliency model research. The dataset has 1003 color images. A non-intrusive video-based eye tracker with an angle accuracy of 1° and a sample rate of 240 Hz was used to record the human gaze data when an image is shown to the participants. There are 15 participants in the experiment. They viewed each image for 3 s.

We randomly divided the whole dataset into two subsets – training set (803 images) and testing set (200 images). The model is learned on the training set, testing set is used for evaluation.