A Learned Saliency Map for Eye Detection
Li Xingshan    Yuan Chunwei
Department of Biomedical Engineering
Southeast University
Nanjing, Jiangsu Province,210096
{lixslixs}_{cwy}@seu.edu.cn

Abstract: In this paper, a method of constructing saliency map based on learning method is proposed. The saliency map is constructed by nonlinear combination of features, and the nonlinear function is approached through a feed forward artificial neural network (ANN) whose weight is trained by examples. The efficiency of the proposed method is verified in the eye detection problem. Experiments show that the learned saliency map may be efficiently used in the eye detection problem.

Keywords: eye detection, visual attention, saliency map, artificial neutral network

I.Introduction
Selecting only a subset of the available sensory information before further detailed processing is crucial for efficient perception.[1] In the visual modality, this selection is frequently implemented by suppressing information outside a spatially circumscribed region of the visual field, the so-called “focus of attention.”(FOA) In primates, the identification of behaviorally relevant objects and the analysis of their spatial relationships involve either rapid, saccadic eye movements or so-called “covert”(i.e., without eye movements) shifts of visual attention. While attention can be controlled in a voluntary manner, it is also attracted in a “bottom-up,” automatic and unconscious manner to conspicuous, or “salient,” visual locations. Focal attention is often thought as a gating mechanism, which selectively allows a certain spatial location and certain types of visual features to reach higher visual processes.[2]

Artificial system may employ this mechanism either. Koch and Ullman introduced the idea of a saliency map to accomplish preattentive selection.[3] Saliency map is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among sites in this map gives rise to a single winning location that corresponds to the most salient object. If this location is subsequently inhibited, the FOA is shifted to the next most salient location, which constitutes the next target, endowing the search process with internal dynamics.

Object detecting simulating visual attention has been attempt by many researchers [2,4,5]. In these approaches, the object detection is divided into two stages: ‘where stage’ and ‘what stage’. The ‘where stage’ seeks where salient objects are, and the ‘what stage’ identifies what the object is. Due to the bottom-up approach and exploring only the salient sites, these methods share the virtue of computing efficiency.[2,4] In these works, the most important and challenging is the constructing saliency map from features.[2]

In this paper, a method of constructing saliency map based on learning method is proposed. The saliency map is seen as nonlinear combination of features, and the nonlinear function is approached through a feed forward artificial neutral network (ANN) whose weight is learned through examples. The effect of the proposed method is verified in the eye detection problem.
II The Learned Saliency Map

A central yet not thoroughly studied problem, both in biological and artificial systems related to attention, is that of combining multi-scale feature maps, from different visual modalities with unrelated dynamic ranges (such as color and motion), into a unique saliency map. There are three combination strategies have been employed: (1) Simple summation after scaling to a fixed dynamic range; (2) linear combination with weights learned, for each image database, by supervised additive training; (3) non-linear combination of features. [5] Because the relation between saliency and feature is very complex, linear combination cannot reflect this kind of complex relation. So nonlinear combination should be used to deprive the saliency in complex problems such as object detection. The nonlinear combination may be written as:

\[ \text{Saliency} = f(\text{Features}) \]  

(1)

Where \( f \) is the task related complex nonlinear function. \( \text{Features} \) is a vector of image features.

To approach the object detection problem, the most salient sites in saliency map should correspond to the locations of the object to be detected. Therefore, the function \( f \) should contain information of the object to be detected. One way of introducing knowledge of interested object into \( f \) is learning through examples. Jeffrey Huang[4] deprive the saliency map using a consensus between navigation routines encoded as a finite-state automata exploring the facial landscape and evolved using genetic algorithms. In their approach, the position of the chin is assumed to be known. This is not necessarily true in some cases. And the FSA is based on the routine from chin to eyes, so the FSA in their work depend on far more features than that of eyes. These two points limits the robust of the Huang’s method.

A learned method to approach nonlinear function \( f \) is proposed in this paper. Different to the Huang’s method, the saliency value of one site depends only on the features of a limited surrounding area instead of the whole image.

It is well known that a suitable ANN may approach any function, and can discover underlying statistical regularities of the training example. So, in our approach, a feed forward artificial neural network (ANN) is employed to express the nonlinear function \( f \).

To detect an object in an image, spatial information should be considered. For a site \( o \) in the feature map, the site and the surrounding sites form a Neighboring Site Group (NSG) of \( o \), which contains 3*5 sites. When calculating the saliency value of one site, all features of the NSG are used. The tutor is manually defined to be high value at the locations of interested objects, and low value at others. After trained for several epochs, the ANN may be used to approach the nonlinear function \( f \). In our approach, the Backpropagation (BP) training method is used.

The output of the ANN, which makes up the saliency map of the image, is expected to reach relatively high value at the location of interested objects, and relatively small value at others.

III. Implementation of Eye Detection

Eye detection is an essential task for automated facial recognition system. To design an automated face recognition (AFR) system, one needs to address several related problems: 1) detection of a pattern as a face (in the crowd). 2) detection of facial landmarks, 3) identification of the faces, and 4) analysis of facial expression. Locating the eyes serves, first of all, as an important role in face normalization and thus facilitates further localization of facial landmarks.[6]

There are two major approaches for automated eye location. The first approach, the holistic one, conceptually related to template matching, attempts to locate the eyes using global representations. The second approach for eye detection, the abstractive one, extracts (and measures) discrete local features, while standard pattern recognition techniques are then employed to locate the eyes using these measurements.[4] The method in this paper draws
The eye detection process that being used in this paper simulates the visual attention process of primitives. The task is divided into two stages: ‘where stage’ and ‘what stage’. The ‘where stage’ seeks where salient objects based on the learned saliency map, and the ‘what stage’ identifies what the object is. Only the first stage is considered in this paper to verify the effect of the learned saliency map.

The process is introduced as following, which is also shown in figure 1.

1) Feature Maps: The input consists of detected face images whose resolution are 200*180 using 8 bit gray levels. Three feature maps corresponding to mean, standard deviation (std), and entropy are then computed over 6*6 non overlapping windows, and then compressed to yield three 34*30 feature maps.

2) The Trained Saliency Map: After the feature extracting process, the images are transformed into three feature maps of 34*30 sites. To deprive a saliency map, for each site, the features of the NSG are used. Facial images of different individual with different expression are selected as the training data set. For every training image, the eye locations are manually extracted, sampled to the size of 34*30, which is known as the eye mask. The eye mask is then filtered with an 3*3 average filter, which will makes the center of the eye has the most high value. The processed eye mask acts as the tutor to train the ANN. After trained, the ANN may be used to construct the saliency map of an image. The features of each site’s NSG are fed to the trained ANN, the output of the ANN is the saliency value of that site.

3) Transfer of Focus of Attention(FOA): The FOA is an area of interested that will be future processed. The FOA, whose size is 5*9 in this paper, is located to the location of the most salient site in saliency map. Then eye identifying methods (which will be explained in next segment) are used to detect if an eye present in an FOA. After one FOA being...
processed, the attention will be transferred to the next salient FOA. The process stops when both eyes have been detected.

4) Eye identification: When focus on one FOA, traditional methods such as model matching or other methods may be used to detect if an eye appear in the FOA. This stage is beyond the range of this paper. For the purpose of verifying the effect of the learned saliency map, a manual eye identifying method is employed in this paper.

IV. Experiments and Results

The experiment material comes from the Vision Group of Essex University Face Database. [7] The test set contains face images of 18 male and female subjects, with image resolution 200*180 pixels (portrait format). A sequence of 20 images per individual was taken, using a fixed camera. During the sequence the subject moves his/her head and makes grimaces which get more extreme towards the end of the sequence.

The images are divided into two groups. In the first group with 10 individual, two images of each individual are randomly selected as training samples, and all the other images act as testing images, which build up the first test set. In the second group with 8 individuals, all images act as testing images, which build up the second test set.

Table 1 is a statistical result of the numbers of transfer of focus before both of the two eye locations have been focused. Comparing the two results with different testing data set, we may find that they make no obvious different. This shows that this method is robust to facial image of different individuals.

In the following stage for eye detection, it needs only to explore the FOAs. The average number of FOAs before both eyes detected is fairly small, say, in this experiment, only 2.3438 in the second test set. The site number in a FOA is 35, so average 82 sites needs to be explored instead of the whole image, whose site number is 1020.

<table>
<thead>
<tr>
<th>data set</th>
<th>mean</th>
<th>std</th>
<th>maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>first test set</td>
<td>2.1800</td>
<td>0.5282</td>
<td>7</td>
</tr>
<tr>
<td>second test set</td>
<td>2.3438</td>
<td>0.7443</td>
<td>6</td>
</tr>
</tbody>
</table>

V. Conclusion

By simulating primitive’s visual attention, a saliency map based object detection method is employed in the eye detection problem. The information of the interested object may be represented in the nonlinear function \( f \), which may be approach by a feed forward artificial neural network whose weights are learned through training samples. Experiments show that the learned saliency map may be efficiently used in the eye detection problem.

Though only be tested in the eye detection problem, it is believed that the saliency map based object detection method may be used in other object detection problems.

VI Acknowledgments

This work is financially supported by NSFC (No. 69831010). And thanks to the Vision Group of Essex University for providing Face Database.

References:

Intelligence, vol. 78, no. 1-2, pp. 507-545, Oct. 1995


