A Biologically-inspired Top-down Learning Model Based on Visual Attention

Nong Sang, Longsheng Wei and Yuehuan Wang
Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China
nsang@hust.edu.cn, weilongsheng@163.com, yuehwang@mail.hust.edu.cn

Abstract

A biologically-inspired top-down learning model based on visual attention is proposed in this paper. Low-level visual features are extracted from learning object itself and do not depend on the background information. All the features are expressed as a feature vector, which is looked as a random variable following a normal distribution. So every learning object is represented as the mean and standard deviation. All the learning objects are combined as an object class, which is represented as class’s mean and class’s standard deviation stored in long-term memory (LTM). Then the learned knowledge is used to find the similar location in an attended image. Experimental results indicate that: when the attended object doesn’t always appear in the background similar to that in the learning objects or their combinations change hugely between learning images and attended images, our model is excellent to other two top-down visual attention models.

1. Introduction

The visual system requires attention and guidance of that attention because the eyes provide the central nervous system with more information than it can process. Attention has been classified into bottom-up attention and top-down attention. The top-down attention is more complex to model because it needs to represent object in LTM [1] and uses the memory to detect likely object in attended scenes.

The most popular bottom-up computational model is proposed by L. Itti et al. [2], in which salience according to primitive features such as intensity, orientation and color are computed independently. There are also many top-down visual attention models [3-5]. Well-known models include Visual Object detection with a CompUtational attention System (VOCUS) [3] and the model proposed by NavalPakkam et al. [4]. Top-down attention of VOCUS takes the rate between object and its background as the weight of feature map. The top-down component proposed by Naval-Pakkam uses statistical knowledge to obtain the weight of feature map by maximizing signal-to-noise ratio of object and its background. The performances of these two top-down approaches are influenced by object background, so their effects aren’t very well when the combinations of background and object change hugely.

In order to address the above problem, we propose an attention learning model, which just uses object itself information and does not use its background information. Our model is shown in Figure 1.

![Figure 1. Our model: Given a task such as “find a red cup in the scene”. Firstly, all object’s features are extracted from learning images. Secondly, these features are used to compare the similarity in the input image and form multi-scale similarity maps by priming the desired features, then all the similarity maps are combined into a top-down salience map. Thirdly, top-down map is multiplied by bot-](image-url)
tom-up map and form a final salience map, which is guided attention to likely target locations.

2. Object representation

In this part, basic visual features are extracted from object itself and do not depend on the background information. All the learning objects are combined as an object class, which is represented as class’s mean and class’s standard deviation.

2.1 Feature extraction

For every learning object, eleven low-level visual features including three color features, two intensity features, four orientation features and two texture features are extracted in this passage. We represent each feature as mean and standard deviation \( (\mu_{ij}, \sigma_{ij}) \) (i denotes the i-th learning object, j denotes the j-th feature). Three color features are red, green and blue three channels of the object image. Intensity feature includes intensity on (light-on-dark) and intensity off (dark-on-light). We convert the color image into gray-scale image to obtain an intensity image and let center/surround contrast be intensity on, surround/center contrast be intensity off. Four orientation features \((0^\circ,45^\circ,90^\circ,135^\circ)\) are computed by Gabor filters detecting bar-like features according to a specified orientation. We let the scale of Gabor filters equal to the scale of learning object in this paper. For texture feature, we consider local binary pattern (LBP) [7], which describes the local spatial structure of an image and has been widely used in explaining human perception of textures. Two LBP operators, which are illustrated in Figure 2, are used in this paper. One is original LBP operator and the other is extended LBP operator with a circular neighborhood of different radius size.

![Figure 2. Left: The original LBP operator; Right: The extended LBP operator.](image)

2.2 Learning of object representation

We combine the different objects in the same class obtained above to form a more stable, general representation of an object class that is robust to noise and store the object class in the visual LTM [5]. The following rules are used for combination of several target objects to form a general representation of the object class. Let \( L_i \) be the event that the i-th learning target object occurs, where \( i \in 1,2,\cdots,n \).

Let \( O \) be the random variable denoting an observation and \( O = o \) be the event that the value \( o \) is observed. \( P(O = o | L_i) \) refers to the conditional density, i.e., the probability of observing \( O = o \) given that the i-th object has occurred. Let \( P(O = o | L_i) \) follow a normal distribution \( N(\mu_i, \Sigma_i) \) where \( \mu_i = (\mu_{i,1}, \mu_{i,2}, \cdots, \mu_{i,11})^T \), i.e., a vector of the mean feature values, and \( \Sigma_i \) is the covariance matrix. Due to our assumption that the different features are independent, the covariance matrix reduces to a diagonal matrix, whose diagonal entries equal the variance in feature values, represented as \( \sigma_j^2 = (\sigma_{j,1}^2, \sigma_{j,2}^2, \cdots, \sigma_{j,11}^2)^T \).

Let \( L \) be the event that the learning object class occurs. We define \( L \) as follows:
\[
L = \bigcup_i L_i
\]
In other words, an observation is said to belong to the object class if and only if it belongs to any of the object. Then, we consider the distribution of \( O \mid L \) and obtain the as following:
\[
P(O = o \mid L) = P(O = o \mid \bigcup_i L_i) / P(\bigcup_i L_i)
\]
\[
= \sum_i P(L_i, O = o) / \sum_i P(L_i)
\]
\[
= \sum_i P(O = o \mid L_i) P(L_i) / \sum_i P(L_i)
\]
\[
= \sum_i P(O = o \mid L_i) w_i
\]
where
\[
w_i = P(L_i) / \sum_j P(L_j) = \frac{1}{n}
\]
\[
\mu = \mathbb{E}(O \mid L) = \sum_i w_i \mu_i
\]
\[
\sigma^2 = \mathbb{E}((O \mid L)^2) - (\mathbb{E}(O \mid L))^2 = \sum_i w_i (\sigma_j^2 + \mu_j^2) - \mu^2
\]
In general, \( O \mid L \) has a multi-modal distribution. But as a first approximation and to achieve recursion in our implementation, we consider only up to the second moment and approximate this multi-modal distribution by a normal distribution \( N(\mu, \sigma^2) \).

By processing several learning objects at different poses and sizes in the same class, we learn the representation of the learning objects and combine them to form a representation of the object class.

\[3737\]
3. Attentional selection

For a given attention image, we obtain a top-down visual salience map according to the target object information represented in section two. Then, top-down and bottom-up salience map are fused into a final salience map.

3.1 Top-down salience map

To detect a specific target object in any scene, we use the learning representation object class stored in LTM to bias the combination of different feature maps to form a top-down salience map.

For a given attention image $I$, we extract eleven feature types maps as described in section two. We create a Gaussian pyramid of learning image $I_i(s)$, where $s \in \{1, 2, 3, 4\}$, according to the size of $I$. In this way, each learning image has four sizes, which equal to one second, one fourth, one eighth and one sixteenth respectively of the size of $I$. Then, we choose the size of learning image as the size of sliding window. For coordinate $(x, y)$, we use the sliding window to across every feature map and calculate the mean $\mu_{i,x,y}(x, y)$ (i denotes the $i$-th feature type, $s$ denotes scale) and standard deviation $\sigma_{i,x,y}(x, y)$ at each window. The mean and standard deviation in every window express the characteristic of likely object, which is thought to be similar to learning objects if they are close to the mean and standard deviation of learning objects. We define the response map of the $i$-th feature, the $s$-th scale is:

$$R_{i,s}(x, y) = \frac{1}{|\mu_{i,s}(x, y) - \mu| \times |\sigma_{i,s}(x, y) - \sigma| + 1}$$  \hspace{1cm} (9)

where $1$ in the denominator avoids the $R_{i,s}(x, y)$ to become unbounded; $\mu$ and $\sigma$ is mean response and standard deviation response to the $i$-th feature type of learning object. Let $N(.)$ be normalization operator and $\otimes$ be point-by-point addition. The feature response map is formed by combining response map of different spatial scales and the same feature:

$$R_{i}(x, y) = \otimes_{i=1}^{4} N(R_{i,s}(x, y))$$  \hspace{1cm} (10)

We use spatial competition function $f$ to combine all features to form a top-down salience map. For details regarding implementation of this feature combination strategy, please see Section 2.4 in [6].

$$S_{td}(x, y) = f\left(\sum_{i} N(R_{i}(x, y))\right)$$  \hspace{1cm} (11)

3.2 Final salience map

A bottom-up salience map indicates how conspicuous every location based merely on the image data itself. The bottom-up salience map part in our proposed model is an extension of the model proposed by Itti et al. [2] since the proposed model considers texture feature.

Top-down map and bottom-up salience map are described above. Final salience map is the point-by-point multiplication between the top-down and the bottom-up salience map. Both maps compete for saliency: the top-down map emphasizing the features of the learned object, the bottom-up map showing regions that are salient because of scene-specific conspicuousities. To make the two maps comparable, $S_{td}$ is normalized in advance to the same range as $S_{bu}$.

$$S = S_{td} \times S_{bu}$$  \hspace{1cm} (12)

4. Experimental results

We apply three schemes of attention mechanism such as VOCUS top-down approach [3], NavalPak’s approach [4] and our proposed approach on same scene images and same learning objects. We have done 50 group experiments and each group experiment is done by these three approaches. There are 24 nature scenes including 20 single object scenes and 4 multi-object scenes. At the same time, there are 26 artificial scenes including 17 single object scenes and 9 multi-object scenes.

Taking a nature scene as example, Figure 3 is learning images and Figure 4 is three experimental results. We first introduce three definitions: hit number, average hit number and detection rate. The hit number on an image for one target is the rank of the focus that hits the target in order of saliency. The average hit number for an image set is the arithmetic mean of the hit numbers of all images. Therefore, the lower the hit number, the better the search performance. If target object is found within first 5 attention foci, then visual attention is regarded as a success. We define detection rate as the percentage of the number of scenes that targets detected within the first 5 attention foci to all the number of scenes.

All object scenes experimental results of three approaches are expressed in Table 1 and Figure 5.

![Figure 3. Learning images](image-url)
Figure 4. Left: VOCUS approach, the object is found in the 8th times; Middle: NavalPakkam’s approach, the object is found in the 6th times; Right: Our proposed approach, the object is found in the 3rd times.

Table 1. All object scenes average hit number and detection rate of three approaches

<table>
<thead>
<tr>
<th></th>
<th>VOCUS top-down</th>
<th>Naval-Pakkam</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average hit number</td>
<td>5.8286</td>
<td>4.6897</td>
<td>3.1449</td>
</tr>
<tr>
<td>Detection rate (%)</td>
<td>43.9024</td>
<td>43.9024</td>
<td>74.3902</td>
</tr>
</tbody>
</table>

Figure 5. All object scenes experimental results of three approaches (x-axis expresses the how many times the target object is found; y-axis expresses the total number of emerged in this time.)

According to above experimental results, artificial scenes experimental results are better than nature scenes experimental results. The reason is that the backgrounds of artificial scenes are simple, but the backgrounds of nature scenes are easy to be influenced by others, such as noise, illumination and clutter. Multi-object scenes experimental results are better than single object scenes experimental results. There are more regions similar to target object in multi-object scenes, so there are more chance to find the target object within first several attention foci. Our proposed model is excellent to the top-down approach of VOCUS and NavalPakkam’s statistical model.

5. Conclusion

We have proposed a biologically-inspired top-down learning model based on visual attention in this paper. Low-level visual features are extracted from object itself and do not depend on the background. All the learning objects are combined as an object class, which is represented as class’s mean and class’s standard deviation stored in LTM. Corresponding features are extracted in the attended image according to the size of learning objects. For each feature, the similarity map is obtained by comparing learning feature and attended feature. All the similarity maps are combined into a top-down salience map, which is multiplied by bottom-up map and form a final salience map. Experimental results indicate that: when the attended object doesn’t always appear in the background similar to that in the learning objects or their combinations change hugely between learning images and attended images, our model is excellent to the top-down approach of VOCUS and NavalPakkam’s statistical model. The model accords with the habit of human vision system and provides a new method for top-down visual attention.

Acknowledgements

This work was supported by the Chinese National 863 Grant No. 2009AA12Z10 and National Natural Science Foundation of China under contracts 60736010.

References