A Reverse Hierarchy Model for Predicting Eye Fixations

Abstract

A number of psychological and physiological evidences suggest that early visual attention works in a coarse-to- fine way, which lays a basis for the reverse hierarchy theory (RHT). This theory states that attention propagates from the top level of the visual hierarchy that processes gist and abstract information of input, to the bottom level that processes local details. Inspired by the theory, we develop a computational model for saliency detection in images. First, the original image is downsampled to different scales to constitute a pyramid. Then, saliency on each layer is obtained by image super-resolution reconstruction from the layer above, which is defined as unpredictability from this coarse-to-fine reconstruction. Finally, saliency on each layer of the pyramid is fused into stochastic fixations through a probabilistic model, where attention initiates from the top layer and propagates downward through the pyramid. Extensive experiments on two standard eye-tracking datasets show that the proposed method can achieve competitive results with state-of-the-art models.

1 Introduction

Human vision system can selectively direct eyes to informative and salient parts of natural scenes. This ability allows adaptive and efficient allocation of limited computational resources to important objects. Though enjoying great potential in various applications of computer vision, predicting eye fixations, however, remains a challenging task. The underlying difficulty inherits from the ambiguous notion of what attracts eye fixations, or what is salient. In fact, the theoretical investigation of visual saliency has aroused enduring controversies. One possible explanation often adopted in the design of saliency detection approaches is the Feature Integration Theory (FIT). According to FIT, attention serves as a mechanism to coherently combine features for the perception of objects. Therefore, starting from , eye fixations are commonly predicted by directly conjoining saliency activations from multiple channels, which can be global and local channels, multiple features and so on.

Anatomical and physiological studies have shown that human visual system is organized hierarchically, which is believed to be advantageous in efficient processing of visual input. Computational studies have shown that hierarchical models (e.g. HMAX, CDBN) are effective for object recognition. Most saliency detection models, however, do not seriously take this into account. An obvious method to fill this gap is to develop hierarchical bottom-up models for saliency detection in the manner of HMAX, CDBN and the like. But there exists theoretical alternatives. The Reverse Hierarchy Theory (RHT) argues that parallel feedforward feature activation acts implicitly at first to construct a coarse gist of the scene, while explicit perception incrementally incorporates fine details via feedback control. This theory potentially has tremendous applications in computer vision including image segmentation, object recognition and scene understanding, however, computational studies are scarce. In this paper, we present an effective model based on RHT for saliency detection, which proves that RHT is helpful at least in this particular computer vision application. As for this application, a more direct evidence for the proposed model refers to a psychophysical study which showed that fixations from low-resolution images could predict fixations on higher-resolution images.

.

Our main idea is to model the coarse-to-fine dynamics of visual perception. We take a simple strategy to construct a visual hierarchy by inputting images at different layers with different scales, obtained by downsampling the original image. The higher layers receive coarser input and lower layers receive finer input. On each layer, saliency is defined as unpredictability in coarse-to-fine reconstruction through image super-resolution. The saliency on each layer is then fused into fixation estimate with a probabilistic model that mimics reverse propagation of attention. Throughout the paper, we call the proposed model a reverse hierarchy model (RHM).

The coarse-to-fine dynamics, however, is not the only property of RHT. In fact, RHT is closely related to the biased competition theory of attention, which claims that attentional competition is biased by either stimulus-driven or task-dependent factors. Our model deals with fixation prediction in the free viewing task, which can be regarded as an implementation of the stimulus-driven bias. In addition, the image pyramid is a very coarse approximation of the highly complex structure of the visual hierarchy in the brain, which only utilizes the fact of increasing receptive field sizes along the hierarchy. Therefore, some closely related concepts to RHT, such as perceptual learning, would not be discussed in the paper.

2 Related Work

The majority of computational attention modeling studies follow the Feature Integration Theory. In particular, the pioneering work by first explored the computational aspect of FIT by searching for center-surround patterns across multiple feature channels and image scales. This method was further extended through integration of color contrast, symmetry, etc. Random Center Surround Saliency adopted a similar center-surround heuristic but with center size and region randomly sampled. introduced a graph-based model that treated feature maps as fully connected nodes, while the nodes communicated according to their dissimilarity and distance in a Markovian way. Saliency was activated as the equilibrium distribution.

Several saliency models adopted a probabilistic approach and modeled the statistics of image features. and Baldi defined saliency as surprise that arised from the divergence of prior and posterior belief. SUN was a Bayesian framework using natural statistics, in which bottom-up saliency was defined as self-information. proposed an attention model based on information maximization of image patches. defined the saliency by computing the Hotelling's T-squared statistics of each multi-scale feature channel. considered saliency in a discriminative setting by defining the KL-divergence between features and class labels.

A special class of saliency detection schemes was frequency-domain methods. proposed a spectral residual method, which defined saliency as irregularities in amplitude information. explored the phase information in the frequency domain with a Quaternion Fourier Transform. Recently, introduced a simple image descriptor, based on which a competitive fast saliency detection algorithm was devised.

Different from our proposal, the conventional practice in fusing saliency at different image scales and feature channels was through linear combination. proposed a model that combined a global saliency model AIM and a local model through linear addition of normalized maps. Some models learned the linear combination weights for feature channels. trained a linear SVM from human eye fixation data to optimally combine the activation of several low-, mid- and high-level features. With a similar idea, adopted a regression-based approach.

Our model is characterized by a top-down flow of information. But it differs from most existing saliency detection models that incorporate top-down components such as in two aspects. First, a biased prior (e.g., context clues, object features, task-related factors) is often needed in those models, serving as the goal of top-down modulation, which is not necessary in our model. Second, hierarchical structure of the visual cortex is not considered in those models, but plays a significant role in our model.

Nevertheless, there were a few preliminary studies trying to make use of the hierarchical structure for saliency detection and attention modeling. The Selective Tuning Model was such a model. It was a biologically plausible neural network that modeled visual attention as a forward winner-takes-all process among units in each visual layer. A recent study used hierarchical structure to combine multi-scale saliency, with a hierarchical inference procedure that enforces the saliency of a region to be consistent across different layers.

3 Saliency from Image Super-Resolution

In this section, a coarse-to-fine saliency model based on image super-resolution is presented. We consider an image at two consecutive scales in an image pyramid: a coarse one I_l and a fine one I_h. Inspired by RHT, we define saliency as details in I_h that are unpredictable from I_l. In the next section, we discuss how to fuse saliency on each layer of the pyramid into fixation estimate.

3.1 Saliency as Unpredictability

Predicting I_h using the information of I_l is closely related to image super-resolution, which has been extensively studied using techniques including Markov random field, example-based learning, compressive sensing, etc. In patch-based representation of images, the problem is to predict a high-resolution $H \times H$ patch $x_h \in I_h$ from its low-resolution $L \times L$ counterpart $x_l \in I_l$. For convenience of notation, we also use x_h and x_l as H^2 and L^2 dimensional vectors, which are computed by reshaping the corresponding patches. Then x_l is obtained by blurring and downsampling x_h:

$$x_l = GBx_h, \tag{1}$$

where B denotes a $H^2 \times H^2$ blurring matrix (throughout the paper a Gaussian matrix is used) and G represents a $L^2 \times H^2$ downsampling matrix. Let z_h denote the reconstructed patch by some method A, which summarizes the best knowledge one can recover from the coarse perception of x_l, via A. The reconstruction error of z_h from x_h, naturally represents the fine-scale information that cannot be recovered. Therefore, we define saliency $S(x_h|x_l)$ as the Normalized Mean Square Error (NMSE):

$$S(x_h|z_h) = \frac{||x_h - z_h||^2}{||x_h||^2} \tag{2}$$

The mean squared error is normalized so that $S(x_h|x_l)$ is robust to variations of the patch energy $||x_h||^2$.

3.2 Coarse-to-Fine Reconstruction

The reconstruction from the coarse scale subject to the constraint (1) is actually not well-defined, since given a low-resolution patch x_l, there exists an infinite number of possible high-resolution patches x_h. To resolve this issue, the basic idea is to incorporate some prior knowledge, which inherits from the properties of natural images. In what follows we discuss several possible reconstruction schemes with increasingly sophisticated prior knowledge.

Linear Reconstruction (LR). Consider a trivial case: the coarse patch $x_l = Bx_h$, is just the blurred version and we do nothing but output $z_h = x_l$. Therefore, no prior is used in this case. Saliency can be computed according to (2). As shown in Fig. 2, this method assigns more saliency to patches containing many high-frequency components like edges and textures.

Bicubic Interpolation (BI). If we reconstruct x_h using bicubic interpolation, then we utilize a smoothness prior in image interpolation. Although this approach concentrates less on edges than the linear reconstruction, its prediction is still far from the ground truth. See Fig. 2.

With LR or BI, the saliency computed in (2) is the normalized l_2-norm of the Laplacian pyramid. In addition, the two techniques can be used to implement the center-surround strategy adopted in some saliency models, e.g. .

Compressive Sensing (CS). We now consider a more sophisticated prior of image structure – sparsity. According to this prior, any patch x_h of a high-resolution image can be sparsely approximated by a linear combination of items in a dictionary D_h:

$$x_h \approx D_h \alpha, \tag{3}$$

for some sparse coefficients α that satisfies $||\alpha||_0 \leq K$ for some small K. Assuming α is sparse, the theory of compressive sensing states that α can be recovered from sufficient measurements $x_l = GBx_h$ by solving the following optimization problem:

$$\min ||\alpha||_0 subjectto ||D_l \alpha - x_l|| < \epsilon, \tag{4}$$

where $D_l = GBD_h$, denotes the blurred and downsampled dictionary D_h, and ϵ is the allowed error tolerance. This is hard to solve, and in practice the following relaxed problem is often solved:

$$\min ||\alpha||_1 subjectto ||D_l \alpha - x_l|| < \epsilon. \tag{5}$$

The coefficients α are then used to reconstruct z_h by

$$z_h = D_h \alpha. \tag{6}$$

Once we have obtained z_h, saliency of the image patch can be computed using (2). Preliminary results in Fig. 2 indicate that the saliency obtained by compressive sensing can largely differ from that obtained by LR and BIL.

The dictionaries D_h and D_l are constructed as follows. For each scale of the image pyramid, we first uniformly sample raw patches $\{d_j\}_{j=1}^n$ of size $H \times H$ ($n > H^2$), and stack them into a high-resolution dictionary $D_h = [d_1, d_2, ..., d_n]$. Then we apply the blurring matrix B and downsampling matrix G to each d_j, to obtain $d_j = GBd_j$. So $D_l = [d_1, d_2, ..., d_n]$ is the collection of corresponding low-resolution patches. The use of overcomplete raw patches for D_h and D_l has been shown effective for image super-resolution.

3.3 Saliency Map

A saliency map M is obtained by collecting patch saliency defined in (2) over the entire image. First, calculate

$$M[i,j] = S(x_h[i,j] | x_l[i,j]), \tag{7}$$

where $x_h[i,j]$ is the patch centered at pixel (i,j) in the image and $x_l[i,j]$ is its low-resolution version. Then M is blurred with a Gaussian filter and normalized to be between $[0,1]$ to yield the final saliency map M. One should not confuse this Gaussian filter with B in Sections 3.1 and 3.2.

4 Reverse Propagation of Saliency

Now, we present a method to transform the saliency maps at different scales into stochastic eye fixations on the original image. Based on RHT, a reverse propagation model is presented, where attention initiates from top level and propagates downward through the hierarchy.

4.1 Generating Fixations

We model attention as random variables $A_0, A_1, ..., A_n$ on saliency maps $M_0, M_1, ..., M_n$, which are ordered in a coarse-to-fine scale hierarchy. Specifically, let $Pr[A_k = (i,j)]$ denote the probability for pixel (i,j) attracting a fixation. To define this probability, we need to consider factors that influence the random variable A_k. First of all, the saliency map M_k is an important factor. Pixels with higher values should receive more fixations. Second, according to RHT, attention starts from M_0, and then gradually propagates down along the hierarchy. Therefore, A_k should also depend on $A_{k-1}, ..., A_0$. For simplicity, we assume that only A_{k-1} has an influence on A_k while $A_{k-2}, ..., A_0$ do not.

Based on these considerations, we define

$$Pr[A_k | M_k, A_{k-1}, ..., A_0] = Pr[A_k | M_k, A_{k-1}], \tag{8}$$

for $k = 1, ..., n$. A log-linear model is used for this conditional probability

$$Pr[A_k = (i,j) | M_k, A_{k-1}] \propto \exp(\eta M_k[i,j] + \lambda L(A_k, A_{k-1})), \tag{9}$$

where $L(A_k, A_{k-1})$ is a spatial coherence term, η and λ are two constants. The spatial coherence term restricts the fixated patches to be close in space. The motivation of introducing this term comes from the fact that the visual system is more likely to amplify the response of neurons that is coherent with initial perception. To compute the term, we first convert the coordinate A_{k-1} into the corresponding coordinate (u,v) in the saliency map just below it, i.e. M_k. Then compute

$$L(A_k, A_{k-1}) = -((i-u)^2 + (j-v)^2). \tag{10}$$

In other words, the farther away a patch x is from A_{k-1}, the less likely it would be attended by A_k. Therefore, for predicting the fixation probability of any patch in the current layer, the model makes a tradeoff between the spatial coherence with previous attention and its current saliency value.

If we do not consider any prior on the top layer, $Pr[A_0]$ depends on the saliency map only

$$Pr[A_0 = (i,j)] \propto \exp(\eta M_0[i,j]). \tag{11}$$

We can then generate fixations via an ancestral sampling procedure from the probability model. Specifically, we first sample fixation A_0 on map M_0 according to (11), and then for $k = 1, 2, ...$ sample A_k on map M_k given A_{k-1} on the coarser scale according to (9). Finally, we collect all samples on the finest scale, and use them as prediction of the eye fixations.

4.2 Incorporating Prior of Fixations

The proposed probabilistic model offers great flexibility for incorporating prior of fixations. This prior can be useful in capturing, for example, the top-down guidance of visual saliency from recognition, or central bias in eye-tracking experiments. To achieve this, we extend the expression of $Pr[A_0]$ as follows:

$$Pr[A_0 = (i,j)] \propto \exp(\eta M_0[i,j] + \theta P[i,j]), \tag{12}$$

where $P[i,j]$ encodes the prior information of pixel (i,j) on the first map M_0 and θ is a weighting parameter.

For example, the central bias can be incorporated into the model by setting $P[i,j] = -[(i-c_x)^2 + (j-c_y)^2]$, where (c_x, c_y) denotes the map center.

5 Experiments

5.1 Experiment Settings

Datasets. The performance of the proposed reverse hierarchy model (RHM) was evaluated on two human eye-tracking datasets. One was the TORONTO dataset. It contained 120 indoor and outdoor color images as well as fixation data from 20 subjects. The other was the MIT dataset, which contained 1003 images collected from Flicker and LabelMe. The fixation data was obtained from 15 subjects.

Parameters. The raw image I in RGB representation was downsampled by factors of 27, 9, 3 to construct a coarse-to-fine image pyramid. The patch size for super-resolution was set as 9×9 on each layer. To construct corresponding coarse patches, we used Gaussian blurring filter B ($\sigma = 3$) and downsampling operator G with a factor of 3. A total of 1000 image patches were randomly sampled from all images at the current scale to construct the dictionary D_h, which is then blurred and downsampled to build D_l.

In some experiments, we included a center bias in the model. This is achieved by switching θ from 0 to 1 in (12).

Note that the reverse propagation described in (8)-(11) is a stochastic sampling procedure and we need to generate a large number of fixations to ensure unbiased sampling. We found that 20000 points on each image were enough to achieve good performance, which was adopted in all experiments. The stochastic points were then blurred with a Gaussian filter to yield the final saliency map. The standard deviation of the Gaussian filter was fixed as 4 pixels on saliency maps, which was about 5

Evaluation metric. Several metrics have been used to evaluate the performance of saliency models. We adopted Area Under Curve (AUC), Normalized Scanpath Saliency (NSS) and Similarity (S). Specifically, We used the AUC code from the GBVS toolbox, NSS code from and Similarity code from . Following , we first matched the histogram of the saliency map to that of the fixation map to equalize the amount of salient pixels in the map, and then used the matched saliency map for evaluation. Note that AUC was invariant to this histogram matching.

Models for comparison. The proposed model was compared with several state-of-the-art models: Itti Koch, Spectral Residual Methods (SR), Saliency based on Information Maximization (AIM), Graph Based Visual Saliency (GBVS), Image Signature (ImgSig), SUN framework and Adaptive Whitening Saliency (AWS). The implementation of these models were based on publicly available codes/software. Among these models, GBVS, ImgSig and AWS usually performed better than the others.

Inspired by the center bias, we included a Center model as a baseline, which was simply a Gaussian function with mean at the center of the image and standard deviation being 1/4 of the image width. This simple model was also combined with other saliency detection models to account for the center bias, which could boost accuracy of fixation prediction. Following , this was achieved by multiplying the center model with the saliency maps obtained by these models in a point-wise manner.

5.2 Results

First, we compared different super-resolution techniques (LR, BI and CS) for eye fixation prediction. Fig. 5 shows the results of RHM with the three techniques. The CS method significantly outperformed LR and BI. Therefore, sparsity as a prior offers great advantage in discovering salient fine details. We then focused on RHM with CS in subsequent experiments.

Fig. 4 shows some qualitative comparison of the proposed model against existing models. Table 5 shows quantitative results under three metrics. As we can see, no single model could dominate others under all three metrics. However, in most cases (including both "with" and "without center" settings), the RHM outperformed the current state-of-the-art models. This demonstrated the reverse hierarchy theory as a promising way to predict human eye fixations.

5.3 Contributions of Individual Components

The RHM consists of two components: coarse-to-fine reconstruction (especially compressive sensing) and reverse propagation. Although the two components integrated together showed promising results, the contribution of each component to the performance is unclear. This is discussed as follows.

Compressive sensing. To identify the role of compressive sensing, we substituted it with other saliency models. Specifically, we replaced the saliency maps obtained from coarse-to-fine reconstruction by the saliency maps obtained by existing models. The models designed to work on a single scale, including SR, AIM, SUN, were applied to images of different scales to obtain multiple saliency maps. For multi-scale models such as Itti Koch, we use their intermediate single-scale results.

Notice that blurring with a Gaussian filter is a necessary step in our model to obtain a smooth saliency map from stochastic fixations. Previous results have shown that blurring improved the performance of saliency models. For the sake of fairness, we also tested the models with the same amount of blurring (the sigma of Gaussian) used in RHM. Fig. 6 shows the results on the TORONTO dataset.

The reverse propagation procedure improved the AUC of these models. However, their performance is still behind RHM. Therefore, compressive sensing is a critical component in the RHM.

Reverse propagation. To investigate the effect of reverse propagation, we substituted it with linear combination of saliency maps, which is widely adopted in literature. Table 2 shows the results. The linear combination produced an AUC between the best and worst that a single saliency map could achieve. However, RHM outperformed the best single-map performance. Therefore, through reverse propagation, RHM could integrate complementary information in each map for better prediction.

6 Conclusion and Future Work

In this paper, we present a novel reverse hierarchy model for predicting eye fixations based on a psychological theory, reverse hierarch theory (RHT). Saliency is defined as unpredictability from coarse-to-fine image reconstruction, which is achieved by image super-resolution. Then a stochastic fixation model is presented, which propagates saliency from from the top layer to the bottom layer to generate 01xation esti- mate. Experiments on two benchmark eye-tracking datasets demonstrate the effectiveness of the model.

This work could be extended in several ways. First, it is worth exploring whether there exist better super- resolution techniques than compressive sensing for the pro- posed framework. Second, it is worth exploring if the ideas presented in the paper can be applied to a hierarchical struc- ture consisting of different level of features, which play a signi01cant role in the top-down modulation as suggested by RHT. Finally, in view of the similar hierarchical structure used in this study for saliency detection and other studies for object recognition, it would be interesting to devise a uni01ed model for both tasks.