A TWO-STAGE APPROACH TO SALIENCY DETECTION IN IMAGES

Zheshen Wang, Baoxin Li

Dept. of Computer Science & Engineering
Arizona State University, Tempe, AZ 85287, USA

ABSTRACT

Researches in psychology, perception and related fields show that there may be a two-stage process involved in human vision. In this paper, we propose an approach by following a two-stage framework for saliency detection. In the first stage, we extend an existing spectrum residual model for better locating visual pop-outs, while in the second stage we make use of coherence based propagation for further refinement of the results from the first step. For evaluation of the proposed approach, 300 images with diverse contents were manually and accurately labeled. Experiments show that our approach achieves much better performance than that from the existing state-of-art.

Index Terms— *Image Processing, Image Analysis, Object detection, Pattern Recognition*

1. INTRODUCTION

Studies [1] in psychology and cognition fields have found that, when looking at an image, our visual system would first quickly focus on one or several "interesting" regions of the image before further exploring the contents. These regions are often called salient regions. Visual saliency is in general too subjective a concept to be strictly defined since it is closely related to the viewer's personal preferences, experiences, intentions (e.g., a specific searching task), and etc. Nevertheless, there exist some simple principles that underpin the process of selecting salient regions. For example, when one is shown the image in Figure 1 without being given any instruction, in general the attention will be immediately caught by the bar in column 2 and row 3, since its orientation is quite different from others. In this paper, our study is focused on the detection of this type of saliency that is task, experience and preference independent.

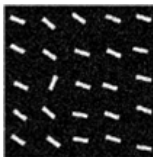


Figure 1.

Many efforts have been devoted to saliency detection [3-7]. Most task independent research follows a bottom-up framework [3] and is often based on searching for regions with maximum local contrast of color, intensity, orientation, etc [5,6]. Some recent work started to seek regional and global features, such as the contrast of region histograms and spatial distributions of colors [4]. [7] proposed a new spectrum residual method which is able to locate visual "pop-outs" very quickly through capturing "noise" in the logarithmic magnitude-frequency curve of a given image. This idea is very simple and seemingly effective. However, it suffers from several drawbacks when it is directly used for saliency detection.

In this paper, we propose a two-stage approach for saliency detection, which is inspired by the typical two-stage processing in human visual system. In the first stage, we extend the spectrum residual method [7] by introducing an automatic channel selection module and a decision reversal module. In the second stage, we propagate the potentially incomplete salient regions based on their similarity and proximity, which are among the basic Gestalt grouping principles in visual perception. To evaluate the performance of our approach, we accurately labeled 300 images from the image database of [4] as the ground-truth and calculated the average precision, recall and F-measure of our approach and compared with other methods. Results show that our method achieves much better performance than the existing approaches, suggesting that our two-stage approach is a promising model for saliency detection.

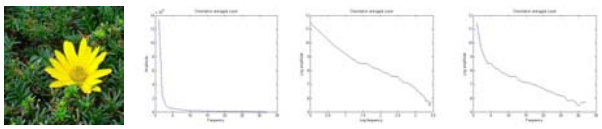
2. TWO STAGES OF VISUAL PROCESSING

Many researches in psychology, perception and cognition, and neuroscience indicate that the human visual system follow two sequential stages in visual perception: The first stage, called pre-attentive stage, processes all the information available fast but coarsely, while the second stage (named focused attention stage) processes only part of the input information with more intensive efforts of exploration [8, 9].

In our approach, we attempt to follow the same procedure: In the first stage, we use a method based on the spectrum residual model [7] to quickly locate the visual pop-outs from the entire image. In this stage, the algorithm extracts only coarse "unusual" regions. In the second stage, our approach takes Gestalt features—similarity and continuity into consideration and propagates the result from the first stage based on local coherence so as to capture some details that are missed in the first stage. These two stages are described in detail in Sections 3 and 4 respectively.

3. COARSE SALIENT REGIION DETECTION BASED ON THE SPECTRUM RESIDUAL MODEL

It has been shown that natural images have a spectrum with the amplitude $A(f)$ obeying the so-called *1/f* law [11,12], as illustrated in Figure 2-a,b. On a log-log scale, the frequency-orientation averaged amplitude curve lies approximately on a straight line (Figure 2-(c)). When represented in frequency-log(amplitude) scale, it becomes a curve with similar trends for any natural images (Figure2-(d)).



(a) (b) (c) (d)

Figure 2. An example of an image and its Fourier spectrum curves: (a) Original image; (b) frequency-orientation averaged amplitude curve; c. (b) on log-log scale; d. frequency-log(orientation averaged amplitude) curve;

It was argued in [7] that the local spiky parts in the frequency-amplitude curve correspond to those sharp changes in the original image, which may be used for saliency detection. Following the basic steps from [7], we first compute the Fourier Transform of an input image, and then take the difference between the transform and its smoothed version (amplitude only). The residual is used in conjunction with the original phase to compute an inverse Fourier transform, which is smoothed with a Gaussian filter to obtain a saliency map. While this simple procedure has been reported to give better performance than Itti's classic framework [3], there are two significant issues that have not been addressed. Firstly, when applying the method to color images, one faces the problem of how to use the spectrum residual model properly. For example, would it be sufficient to process only the luminance channel, or is it effective to process the R,G,B channels separately? Secondly, depending on the contents of the image, often the spectrum residuals may actual correspond to the background rather than to the (foreground) salient region. We term this as the "saliency reversal" problem. In the following two subsections, we propose two techniques to address these issues respectively.

3.1. Channel Selection

If a region in an image is deemed as a salient region, at least one of its visual channels should be different from the rest. In this study we consider the HSV color space. If we only capture one of the channels and the actual contrast mainly resides in other channels, the algorithm would fail. Figure 3 shows an example: (a) is the original image. If we do saliency detection in the hue channel, we can get (c) in which the salient red region is preserved (The brighter, the more salient). If the salient detection is only from the luminance channel, the result would almost lose its target, as in (d). (Our experiments have shown that for the hue or the saturation channel of natural images, the frequency-log(amplitude) curves have similar nice properties of that from the luminance channel.)

(a) (b) (c) (d)

Figure 3. An example of channel selection: (a) Original image; (b) Gray (Intensity) image of (a); (c) Saliency map for hue channel; (d) Saliency map for intensity channel.

To this end, we designed an automatic technique to select the most effective channel. We first compute the saliency maps for each of the three channels: intensity, hue, and saturation. We can get three saliency maps in which each pixel has a saliency score within [0,1]. Then we use k-means clustering for binary clustering, the initial centroids are set as 0 (for non-salient) and 1 (for salient region). We further select the saliency map with the largest distance between two centroids.

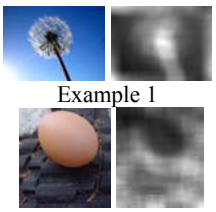
$$EffectiveC\ hannel = \arg\max_{x}(|centroid\ 1_x - centroid\ 2_x|)$$

where x can be the hue, the saturation or the intensity channel. This is based on the assumption that the contrast between the salient and non-salient regions should be maximized before saliency can be detected effectively. This strategy also effectively help us to avoid the problem of selecting optimal thresholds for segmenting salient and non-salient regions, since segmentation is implicitly done by the clustering step. In our experiment using 300 images, in 47.7% of the cases the hue channel is selected, with 33.7% for saturation and 18.7% for intensity.

There may be a potential problem in applying the spectrum residual model to the hue channel: the hue is typically represented as angles; when applying Fourier transform to the angles, we need to set a cutoff point as the zero point. No matter where we set the cutoff, colors in the two sides of the cutoff would have the largest difference (in terms of their angular values) although they are very close colors to each other. This might suggest a potential problem to the spectrum residual model, since two close colors in a smooth region would have extremely different values and thus depict sharp changes. However, we found that if the problem appears within a salient region, it would even somewhat help saliency detection, since an original smooth part of the salient region may pop out more due to the extra frequency introduced by the cutoff of the hue band. Although, if the cutoff point lies in the non-salient region, it may generate false target, it is unlikely for both the salient and non-salient regions to share a lot of common colors, and thus again this situation would not cause too big a problem.

3.2. Saliency Reversal

As discussed earlier, saliency is distinguished by contrast of visual properties. There are two basic cases of contrast patterns: smooth background with cluttered salient region (Figure 4-Example1) and smooth salient region with clutter background (Figure 4-Example2). Unfortunately, the spectrum residual model is only useful for cluttered regions but not for real salient regions. In cases as Figure 4-Example1, they happen to coincide. However, in cases like Example2, they are different. To deal with the latter case, we use the following technique: we reverse the decision based on the spatial distribution of salient pixels in the raw saliency map. We calculate the spatial variance as follows:



Example 1

Example 2

Figure 4. Two cases of salient region, Example1 and 2 left: original image, right: raw saliency map.

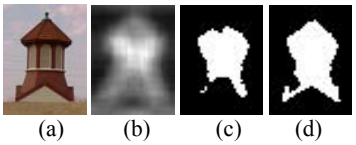
$$\mathrm{var}(R) = \frac{\sum_{i \in R} \sqrt{(r_i - \bar{r}_i)^2 + (c_i - \bar{c}_i)^2}}{size(R)},$$

$$\mathrm{Inverse} = \begin{cases} Yes, \mathrm{var}(background) < \beta \times \mathrm{var}(rawsalientregion) \\ No, otherwise \end{cases}$$

Here, R can be a raw salient region or the background; i is a pixel in R; r_i and c_i are row and column coordinates respectively; $\bar{r}_i$ and $\bar{c}_i$ denote the average row and column coordinates of all pixels in R; $size(R)$ returns the total pixel number in R. β is a threshold in $(0,1]$. In all the experiments on this paper, β is fixed to be 0.85, which works well across all the images. This technique would inverse the raw saliency map when the variance of the original background part is much smaller than that of the raw salient region. Experiments show that this strategy is quit effective in dealing with the problems illustrated earlier.

4. SALIENCY MAP REFINEMENT BASED ON GESTALT GROUPING PRINCIPLES

After the first stage, we can only get the rough locations and regions of saliency. For example, in Figure 5-c the top of the tower is missed in the binary raw saliency map. Naturally, it is desired to have the entire tower detected as a single salient entity. While the top of the tower is not as conspicuous as the middle part, a second stage of the visual processing would group it with the body of the tower. Gestalt refers to theories of visual perception which attempt to describe how people tend to organize visual elements into groups or unified entities when certain principles are applied [10].



(a) (b) (c) (d)

Figure 5 An example of partially detected: (a) Original image; (b) Raw saliency map; (c) Binary raw saliency map; (d) Saliency map after propagation.

Main Gestalt grouping principles include similarity, continuation, closure and proximity. These principles describe visual coherencies from different aspects. Some existing work has developed computable descriptors for each principle [13]. Since our test images are highly diverse and thus some of the principles may not apply (e.g., closure), we experimented with only the most general principles—similarity and proximity, and designed the following method to further process the results from the first step.

In order to retrieve missing regions that should form a unified entity with the extracted regions from the saliency map, we design the following propagation strategy based on similarity and proximity.

(1) Select the largest saliency component in the raw saliency map. (We assume that we are only looking for the largest salient region.). Let's call it "SR".

(2) Divide "SR" into blocks. Find the most representative block "RB", for each block "x" and "y" in "SR", we defined "RB" as the block in "SR" which has the most similar blocks in "SR":

$$RB = \arg\max_{x \in SR} (\underset{y \in SR}{CountNum}(dist(x,y) < threshold)).$$

(3) Find all the neighboring blocks "NB"of "SR" (We use 8 neighbors here.) and compare each of them with "RB". If $dist(NB, RB) < threshold$, add this block to "SR".

(4) Repeat (3), until no new blocks can be added to "SR".

(5) Output "SR" as the propagated saliency region.

We measure the distance between two blocks in HSV color space:

$$dist(x,y) = \sqrt{(H_x - H_y)^2 + (S_x - S_y)^2 + (V_x - V_y)^2}$$

A_i ($A = H, S, V$; $i = x, y$) denotes the average value of channel A in block i.

5. RESULTS, EVALUATION AND ANALYSIS

In our experiment we use the first 300 images from [4] and manually labeled the salient regions accurate-to-contour. We calculate average precisions, recalls and F-measure and further compare our results with Itti's bottom-up framework (Codes downloaded from http://www.saliencytoolbox.net) and intensity only spectrum residual model.

5.1. Image Database, Ground-truth and Performance Measurements

[4] established a huge image database for saliency detection and supplied the ground-truth based on bounding boxes. Such bounding-box-based ground-truth is far from accurate. As illustrated in Figure 6, apparently, result (b) is much more precise than (c), however, they may have very close precisions based on the bounding-box-based ground-truth (d). If we use ground-truth (e), the difference between (b) and (c) would be obvious. More accurate ground-truth leads

(a) (b) (c) (d) (e)

Figure 6. An example of result and different ground-truth: (a) Original image; (b) Our result; (c) An assumed result; (d) Ground-truth of [4]; (e) Our ground-truth;

to more reliable evaluations. For this consideration, we have labeled 300 images accurate-to-contour manually for our evaluations.

With a ground-truth saliency map G, for any detected salient region mask A, we use following measurements:

$$\Pr ecision = \sum_x g_x a_x / \sum_x a_x \,, \quad \text{Re}call = \sum_x g_x a_x / \sum_x g_x$$

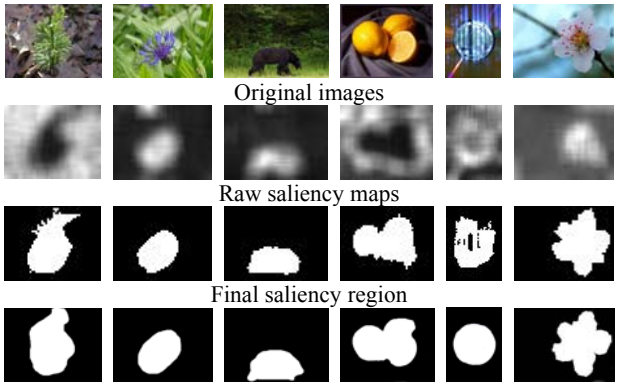
F-measure is the weighted harmonic mean of precision and recall, with a non-negative α :

$$F_\alpha = \frac{(1+\alpha) \times \Pr ecision \times \text{Re}call}{\alpha \times \Pr ecision + \text{Re}call}$$

where α is set to be 0.5 as done usually. If both the precision and the recall are zero, we simply set F_α to zero.

5.2 Results and Comparisons

Figure 7 shows sample results from our approach. We can see that, the final saliency regions not only capture the



Original images

Raw saliency maps

Final saliency region

Ground-truth

Figure 7. Some examples, results and ground-truth

rough location and region of the salient objects, but also roughly keep the contours right. Table 1 shows the statistics

Method	Average Precision	Average Recall	Average F-measure
Itti's bottom-up framework	0.5445	0.1825	0.3052
Original spectrum residual model	0.3013	0.5944	0.3453
Stage I only	0.6230	0.6571	0.5986
Two-stage approach	0.6162	0.7043	0.6122

Table 1 Comparisons of results among different methods

of results from our approach and other methods. For such a challenging image database, classic bottom-up framework has a very poor recall. Intensity-only method has a much higher recall, but its precision is still low. Stage I of our approach doubles the original precision and increases the recall by 6 percentages at the same time and also has a much higher F-measure value, due to the proposed extensions. By adding the second stage, precision decreases a little (less than 1 percentage), but recall and F-measure rise by more than 4 percentages and 1 percentage respectively. In summary, our two-stage approach achieves much better performance than Itti's framework and the original spectrum residual approach.

6. CONCLUSIONS

In this paper, we proposed a two-stage approach for visual saliency detection. For the first state, we extended the spectrum residual model of [7] by introducing automatic channel selection and decision reversal. In the second stage, we develop a coherence propagation strategy based on two basic Gestalt principles. We manually labeled 300 images accurate to contour as the ground-truth for evaluations. Experiments show that our approach performs much better than state-of-art methods, suggesting that this is a promising model for saliency detection.

7. REFERENCES

[1] VF Leavers, "Preattentive computer vision: towards a two-stage computer vision system for the extraction of qualitative descriptors and the cues for focus of attention", *Image and Vision Computing,* 12(9): 583-599, 1994.

[2] Jeremy M. Wolfe, Todd S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?", *Nature Reviews of Neuroscience*, Nature publishing Group, 5: 1-7, 2004.

[3] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention", *Vision Research*, 40(10-12): 1489-1506, 2000.

[4] T. Liu, J. Sun, N. Zheng, X. Tang and H. Shum, "Leanring to Detect A Salient Object", CVPR, 2007.

[5] L. Itti and C. Koch, "Computational Modeling of Visual Attention", *Nature Reviews Neuroscience,* 2(3): 194–203, 2001.

[6] L. Itti, C. Koch, E. Niebur, et al., "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20(11):1254–1259, 1998.

[7] X. Hou, L. Zhang, "Saliency Detection: A Spectral Residual Approach", CVPR, 2007.

[8] A. H. C. van der Heijden, "Two Stages in Visual Information Processing and Visual Perception?", *Visual Congnition*, 3 (4): 325–361, 1996.

[9] Neisser, U. *Cognitive psychology*, Appleton-Century-Crofts, New York, 1967.

[10] Desolneux A., Moisan L. Morel J.-M, "Computational Gestalts and Perception Thresholds", Journal of Physiology-Paris, March 2003, 97(2): 311-324(14), 2003.

[11] D. Ruderman, "The Statistics of Natural Images", *Network:Computation in Neural Systems*, 5(4): 517–548, 1994.

[12] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On Advances in Statistical Modeling of Natural Images", *Journal of Mathematical Imaging and Vision*, 18(1): 17–33, 2003.

[13] S.Bileschi, L.Wolf, "Image representations beyond histograms of gradients: The role of Gestalt descriptors", CVPR, 2007.