โครงงานทางวิศวกรรม

เรื่อง

การปรับปรุงแบบจำลองเชิงคำนวณเพื่อหาความเด่นในภาพเดี่ยว

Improving Computational Model for Visual Saliency on Single Image

โดย

นาย วิรวิชญ์ หรือโอภาศ 5531065821

อาจารย์ที่ปรึกษาโครงงาน ผศ.ดร. ธนารัตน์ ชลิดาพงศ์

ลายมือชื่อ .....

รายงานฉบับนี้เป็นส่วนหนึ่งของวิชาโครงงานวิศวกรรมคอมพิวเตอร์
หลักสูตรวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมคอมพิวเตอร์
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

Abstract

The objective of this project is to improve a computational visual attention model to predict the bottom-up visual saliency on single image. Human receives highly complex and large amount of information every second. Visual attention is a mechanism in human perception, which enables us to selectively process and interpret important parts of what we see. Computational vision systems face the same problems as human. Large amount of available information has to be processed in order to decide the next action in real-time. With fast, premature analysis of visual scene achieved by visual saliency models, the systems can direct their processing power to the areas that are significant to a task at hand.

This project focuses on improving bottom-up visual saliency, the mechanism that requires no higher-knowledge of real world objects and relies purely on low-level features such as color and intensity to guide to interesting regions on given visual scene. Traditional biological-based bottom-up saliency model utilizes three features from early visual pathway in human: Intensity, color and orientation information, which are extracted at the retina, lateral geniculate nucleus (LGN) to the primary visual cortex (V1) respectively. Based on Feature Integration Theory by Anne Treisman, these features compete so that unique features stand out and suggest interesting regions.

Different from the traditional models, this project adopted medium-level features such as corner, line intersection and line ending. Such features usually reside at figure locations, where human pays attention to since those are the essences of a scene. Our experimental results indicate that the proposed model is competitive among other state-of-the-art models. Moreover, we demonstrate that corners or line endings define contour or shape of the figures succinctly and open up possibilities to achieve illusory contours as human perceives.

Keywords *– saliency map; visual attention; corner; bottom-up attention*

Acknowledgements

Firstly, I would like to express my deep gratitude to my project advisor Asst. Prof. Dr. Thanarat Chalidabhongse for all her guidance, encouragement and support throughout my research and study of this work, and also my senior year in undergraduate studies. Her willingness to accept me to sit in on her course and resources has also provided me with sufficient background knowledge on Computer Vision. I would also like to thank all graduate students under her supervision for useful comments that contribute to the development of this work through multiple lab meeting sessions.

I would like to offer my special thanks to Mr. Sangsan Leelhapantu for his corrections and recommendation on both grammatical and structural aspects in my research papers.

My profound gratitude is also extended to my family for all their supports, their love and care for my life.

Wirawit Rueopas

Table of Contents

1. Introduction

1.1 Motivation

Human receives highly complex and large amount of information every second. Visual attention is a mechanism in human perception, which enables us to selectively process and interpret important parts of what we see. This allows us to act and response efficiently in the world. Computational vision systems face the same problems as human. Large amount of available information has to be processed in order to decide the next action in real-time. One way to achieve this is to simulate the attention systems similar to human visual attention.

However, the neural implementation of several visual processes in human especially interactions between different neural systems is still largely unknown. What makes this task intractable is an increasing number of visual phenomena that have been long observed both from experiments and everyday life, such as object and scene recognition, optical illusions, eye movement mechanism and its relationship with visual attention systems, and many more. Moreover, these core functionalities of human vision are tightly connected and working together, and the universal model must be able to cover and explain all, not part of the system.

1.2 Scope

This project focused at *bottom-up visual attention.* It is also called stimulus-driven because the mechanism is mainly based on low-level feature. For example, a single red dot among many green dots on a white paper will have high conspicuity and our eyes are involuntarily attracted to it. Another type of visual attention is *top-down,* or goal-driven. It is based on our internal goal to selectively extract relevant information for current task from visual scene. The example of top-down visual attention is a person looking for an apple among different kinds of fruits. In this case a red oval shape is more salient than other colors or shapes.

Bottom-up processing has been extensively researched, and many well-established computational models are long proposed. Biological-based models usually cover low-level features such as intensity, colors, and orientations. Based on Feature Integration Theory (FIT) by Anne Treisman, these features are extracted independently and can be represented as *feature maps.* They compete for visual saliency by suppress redundancies and promote uniqueness. Finally the features are combined into a final saliency map.

1.3 Contributions

In this work a bottom-up visual saliency model that utilizes corner feature is proposed. The model is based on assumption that figure locations usually reside in areas where there're dense corner, line intersection and line ending information. Note that the term "figure" comes from *figure-ground* organization of Gestalt principles of perceptual grouping. It refers to the process of identifying a figure from the background, which is necessary for object recognition. For example, human usually describe an apple is on a table, rather than a table is behind an apple. This is to emphasize the main content (or a figure) of the visual scene, which is the apple and to denote the table as a background. Sometimes there are no clear decisions that which part is a figure or which part is a ground and they are resolved based on prior knowledge of objects and past experiences of the observer.

Since corners, line intersections and line endings can roughly define contour of figures in visual scene, they can be used to suggest figure locations, which are main contents and areas where people pay attention. The proposed model exploited this fact and has been shown to predict human's eye fixations well on a dataset of natural images compared with other state-of-the-art bottom-up saliency models.

1.4 Outline

The remainders of this document are structured into 5 chapters:

Chapter 2 describes background knowledge consisting of human visual system, visual attention and Gestalt principles of perceptual grouping.

Chapter 3 investigates some of well-known computational visual attention models in the field.

Chapter 4 elaborates the proposed model in details.

Chapter 5 describes related evaluation metrics and shows experimental results of the proposed model with natural image dataset. Other existing models described in Chapter 3 are also compared.

Chapter 6 discusses about the results from Chapter 5 and also possible further improvements on the proposed model.

2. Background on Vision

This chapter describes some of important background knowledge on vision. First an early visual pathway of human is briefly explained (2.1) then follows with an overview of visual attention (2.2). The chapter ends with an introductory on Gestalt principles of perceptual organization.

2.1 Human Visual System

The visual process begins at the retina when the light falls on to the photoreceptors such as rods and cones, which are the only cells that can convert light into neural signal. Rods are extremely sensitive to light. At very low light levels, visual experience is based solely on rod signal, so at dark we often see in gray color. Cones require much brighter light in order to produce a signal. In human, three different types of cones tuned to different wavelengths work together and their signals are then later processed to give us experience of color.

The signal from the photoreceptors then proceeds to bipolar cells, and ganglion cells, and through early visual pathway such as lateral geniculate nucleus (LGN), which receives major input from the retina and relays signals to primary visual cortex (V1) which is the beginning of the visual cortex. A simplified diagram is shown in Fig. 1:

Figure 1. A simplified diagram of early visual pathway in human.

Visual cortex contains multiple parts such as primary visual cortex (V1, and also *striate cortex*) and extra striate areas consist of visual areas two (V2), three (V3), four (V4), and five (V5 or MT or also *the medial temporal area*). V1 responds for extracting basic visual elements such as edge, orientation, direction, spatial frequency and binocularity. It was also probably the most thorough researched area in the Visual cortex. V2 cells are tuned for more complex patterns and may be driven by multiple orientations at different regions in a single receptive field.

Researches for upper layers after V1 and V2 are less precise due to largely complex network between components as shown in Fig. 2.

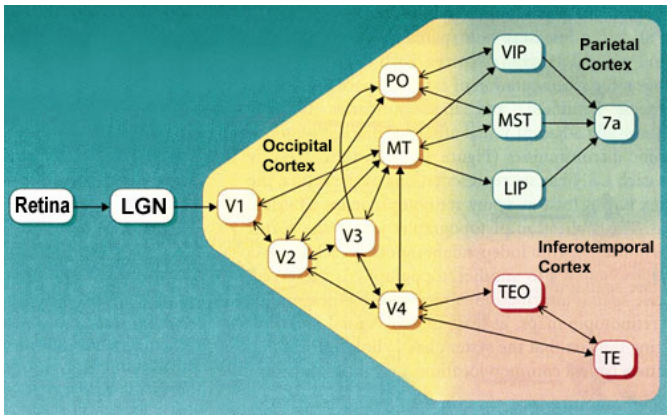

Figure 2. Complex networks inside the visual cortex[1].

Several hypotheses and models were proposed instead. An influential model is *two-streams hypothesis* presented by David Milner and Melvyn A. Goodale in 1982, illustrated in Fig. 3. The model suggests that the visual systems can be divided into 2 major pathways: *dorsal* ("where pathway") and *ventral* ("what pathway") *stream*. The former is involved with processing the object's spatial location and in the guidance of our actions. The latter is associated with object recognition and form representation.

The ventral stream begins at V1, and goes through V2 and V4 to areas of Infero-Temporal cortex (IT). V4 is shown to tune to intermediate complex features such as curve, corner and contour [1] which is essential to object recognition at later stages.

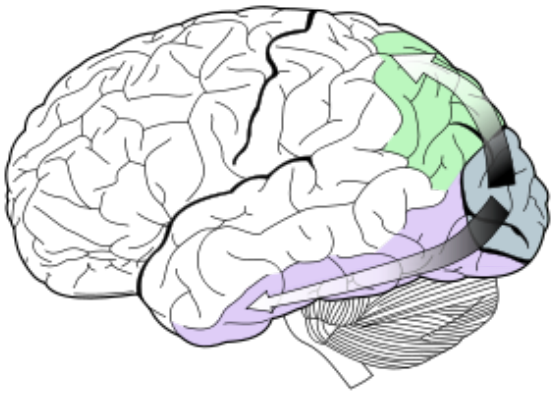

Figure 3. The dorsal and ventral stream in two-streams hypothesis (upper and lower)[2].

[1] Figure from: http://thebrain.mcgill.ca/flash/a/a_02/a_02_cr/a_02_cr_vis/a_02_cr_vis.html
[2] Figure from: https://en.wikipedia.org/wiki/Two-streams_hypothesis

2.2 Human Visual Attention

Formally, attention is the taking possession by the mind one object among several simultaneously ones to deal with. In other words, it is a cognitive process of human for selecting important and relevant information from huge amount of sensory input for further processing.

Attention can be classified as two types:

- Bottom-up processing (stimulus-driven) which concerns only low-level visual stimulus.
- Top-down processing (goal-driven) which concerns observer's current intentions or goals.

Past Development

Feature Integration Theory (FIT) proposed by Anne Treisman and Garry Gelade [2] is one of the most influential theories in the field of visual attention. The theory is illustrated in Figure 4. According to them, the first stage is the pre-attentive stage in which perception occurs automatically, effortlessly, early, and object is analyzed for details such as shape, color, orientation and movement. The second stage of the theory is the focused attention stage, where the individual features of an object are combined in order to perceive the whole object.

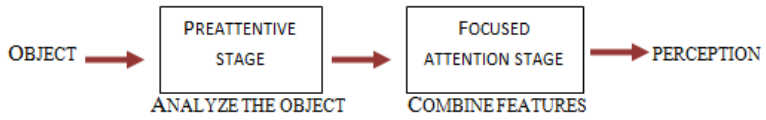

Figure 4. Feature integration theory.[3]

Christof Koch and Shimon Ullman [3] then proposed a feed-forward model to combine these features, and introduced the concept of *saliency map*, which is a map that represents conspicuousness of scene locations. An example of saliency map is shown in Figure 5. Whiter areas indicate higher conspicuous and human may pay attention at it more than other locations. Most computational visual attention models use saliency map as an output of the system.

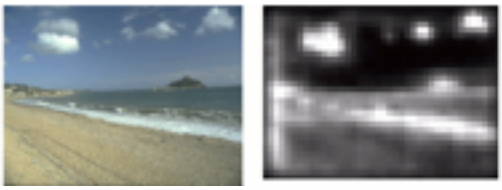

Figure 5. Saliency map that represents conspicuity of an image.[4]

[3] Figure from: https://en.wikipedia.org/wiki/Feature_integration_theory.
[4] Figure from: http://www.scholarpedia.org/article/Saliency_map.

2.3 Gestalts principles of perceptual organization

Gestalt principles are rules of the organization of perceptual scenes. Gestalt means 'shape' or 'form' in German. They are first introduced by Max Wertheimer in 1923 as a theory of mind in psychology school. Traditional scientific methodologies emphasize on dissecting any problems in consideration into parts, find relations between them and put them back together so that the problem is solved. The core idea of Gestalt principles, according to one of the Gestalt psychologist Kurt Koffka, is that the whole is *other* than the sums of its parts. To elaborate, It means that when the mind perceive something, the whole percept is perceived and existed independently on its own and not just an addition from its parts or sub-elements.

In vision, Gestalt laws of grouping often refer to a list of visual phenomenon perceived by the mind from visual elements in the scene due to specific organizations. The foundation of all other principles is figure-ground organization. It is the process of extracting the "figure" from the "background". To put it simply, figure is the main content in the visual scene. An example is shown in Fig. 6, where we instantly perceive the aquamarine oval at the center as a figure.

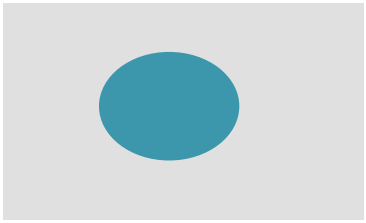

Figure 6. Example of figure-ground.

It may seem trivial to state this as a principle, but it is very important to identify what is figure or what is a ground since sometimes it is not obvious. The well-known example is faces-vase drawing shown in Fig. 7. The drawing can be either faces or a vase. The figure-ground segmentation is influenced by many factors such as color, motion, shape, the prior knowledge or even the current state of mind of an observer. After figure-ground is identified, the Gestalt principles can be applied.

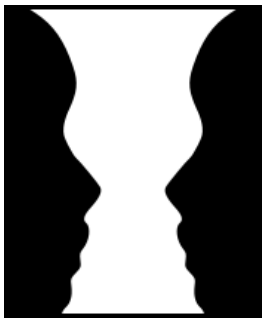

Figure 7. Faces-vase drawing[5].

5 Figure from: https://en.wikipedia.org/wiki/Figure–ground_(perception).

The rest of this section will introduce some of Gestalt principles for perceptual grouping: *proximity*, *similarity* and *continuity*.

Proximity – units near each other tends to be aggregated into groups. Apart from perceiving independent visual units, in (a) it is perceived as the horizontal row. In (b), it is often perceived three groups (12/34/56). It must be note that other arrangements also exist, such as doublet of triples (123/456) but it is hard, if not impossible, to perceive that.

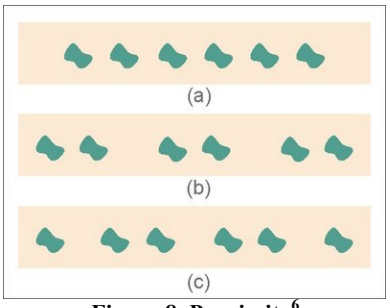

Figure 8. Proximity[6]

Similarity – similar units tend to be aggregated into groups. The viewpoint of certain groupings in proximity can be changed according to the similarity of units. In (a), rather than one horizontal row, it is perceived as three groups.

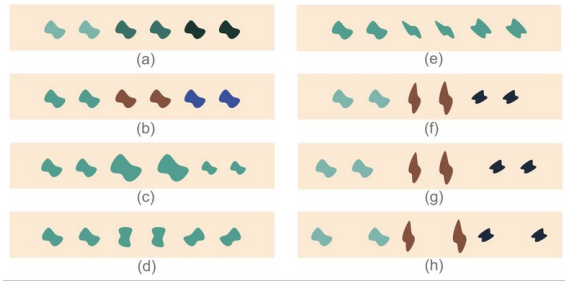

Figure 9. Similarity

Continuity – units tend to be aggregated into perceptual whole if they're aligned with each other. In the below figure, (AXB, XC) is often perceived as two units (curves) as in the right figure, rather than (AXC, XB).

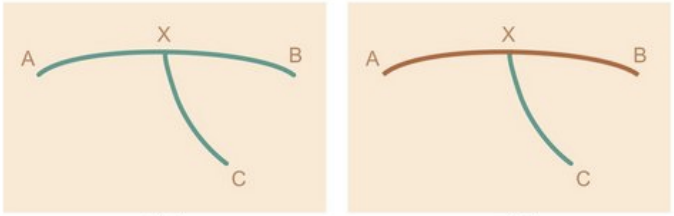

Figure 10. Continuity

[6] Figures in this section are from: http://www.scholarpedia.org/article/Gestalt_principles.

The Gestalt principles give rise to different approaches in computational visual attention models. The main idea is that the visual attention processing of human is not based on pixels of features like colors, intensity and orientation **at each cell** (small units, or pixels) in the visual field, but instead it builds on small cohesive regions called *proto-object*. Some model [4] also includes other Gestalt principles of perceptual organization as capabilities of the model.

3. Previous Work

Most computational visual attention models have the same structure as in Fig. 6. Image or video is input of the system, and the model will compute saliency map as a gray scale image. Some models downsample an input image for efficiency and the final saliency map may be upsampled to the original input size.

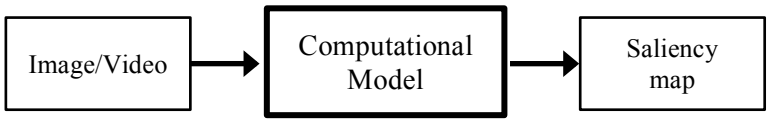

Figure 11. Common structue of computational visual attention model.

In the following sections some of well-known computational models are introduced.

3.1 A model of saliency-based visual attention for rapid scene analysis (1998) [5]

A model proposed by Itti et al. is a classic bottom-up visual attention model, inspired by the feature integration theory by Treisman et al [2]. Most models in the field have adopted it as a bottom-up processing part in their systems. The model is equivalent to early visual processing mechanism as in human, which begins at retina and ends at primary visual cortex. The model is illustrated below in Fig. 9.

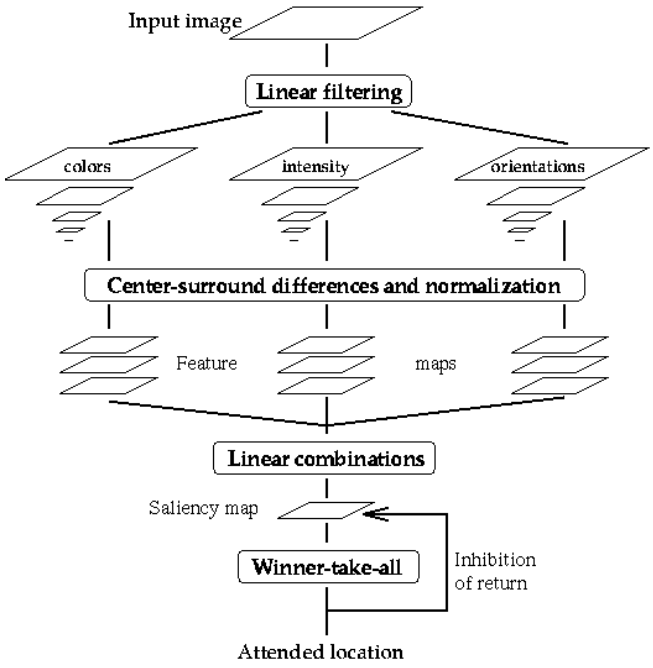

Figure 12. The model of Itti et al. (1998)

The model can be briefly summarized in the following steps:

- Firstly, 9-scale Gaussian pyramid is created from the input image. Three feature types are extracted, containing colors, intensity and orientations. Orientation maps are obtained by filtering intensity pyramids with Gabor filter at 4 different angles, 0°, 45°, 90° and 135°.
- Then, each feature is passed through *center-surround* difference operator, which is across-scale subtraction for representing center-surround mechanism in mammal vision. Across-scale difference is achieved by scaling coarser image up to the same size as the finer one in the feature pyramid, then do point-to-point subtraction. Normalization is also used to bias the highest value in nonlinear way.
- Finally, those maps extracted from each type of feature are combined through across-scale addition (in the same manner as across-scale subtraction) and result in 3 master maps.
- These master maps are then directly combined linearly and the model outputs one saliency map.

This saliency map is then used to predict human eye movement in that scene, typically by selecting the pixel which has the highest saliency value in the map as the first eye location (denoted as *Winner-take-all* in Fig. 9), then a region centered at that location is reduced in their saliency value and to prevent returning to the same point in subsequent eye movements (denoted as *Inhibition of return* in Fig. 9).

3.2 Saliency Based on Information Maximization (2005) [6]

Another bottom-up model proposed by Neil D.B. Bruce and John K. Tsotsos. Unlike the classic bottom-up model by Itti et al., this model uses information theoretic approach, based on the assumption that rarer information will be more interesting and *self-information* is used to represent saliency value. The process is depicted in Fig. 10 below.

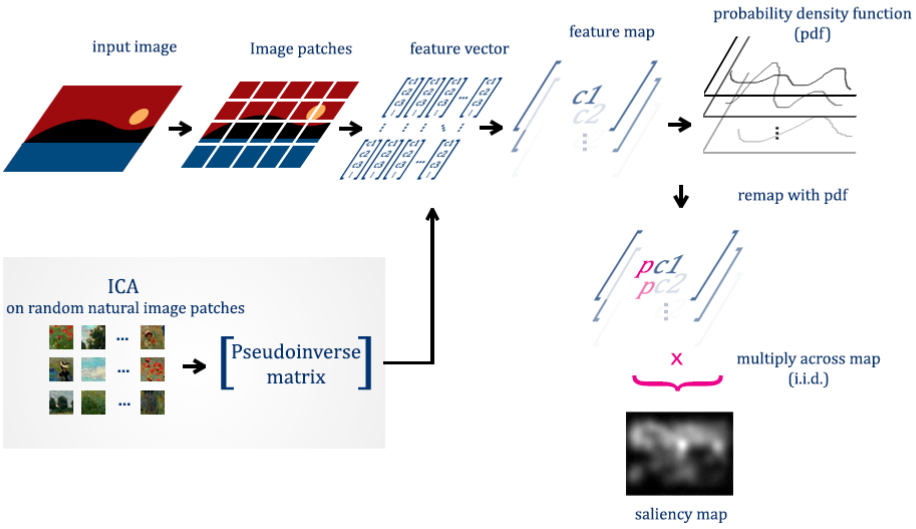

Figure 13. Bruce et al.'s model (2005).

To describe the process briefly, *Independent Component Analysis* (ICA) is applied on random natural image patches to retrieve *pseudo-inverse matrix*, which can be used to convert image patches from input image to 25-dimension feature vectors. Then each dimension is separated into corresponding 25 feature maps and probability density function in each map is obtained either by kernel density estimation or simple histogram.

After that every points on each feature map is remapped with their probability value. Finally, visual saliency of each feature x is defined as:

$$I(x) = -\log\left(p(x_1, x_2, \ldots x_{25})\right)$$

where x_i for i = 1, 2 … 25 is feature vector. With ICA features are assumed to independent to each other, thus the joint probability can be retrieved from:

$$p(x_1, x_2, \ldots x_m) = \prod_{i=1}^{25} p(x_i).$$

It is clearly seen that the saliency value at each input location depends on rareness of features.

3.3 VOCUS: A Visual Attention System of Object Detection and Goal-Directed Search [7]

This model, proposed by Simone Frintop, covers both bottom-up and top-down. By combining both parts, it can search for a target whose features are learnt in advance. It has two modes, learning and searching. Figure 11 illustrates the searching mode.

The bottom-up part of this model is greatly similar to Itti et al.'s model [5]. One difference is that VOCUS differentiates input's intensity map into two parts, namely *on-center* and *off-center*, which reflects closer biological structure in human than the model of Itti et al. This also helps prevent bias between white-on-black and black-on-white stimulus. In addition, color channel, number of scales in Gaussian pyramids, across-scale subtraction and addition are managed differently in VOCUS.

Before doing a visual search task on specified object, the model has to learn about the object first. In learning mode, the model is given an image containing an object desired to learn, with bounding rectangle around that object. Then, it learns about object by comparing their low-level features such as intensity, orientations, and colors against the background area. After that, the object is represented by object's weight vector. The higher the weight of a feature, the more important is that feature.

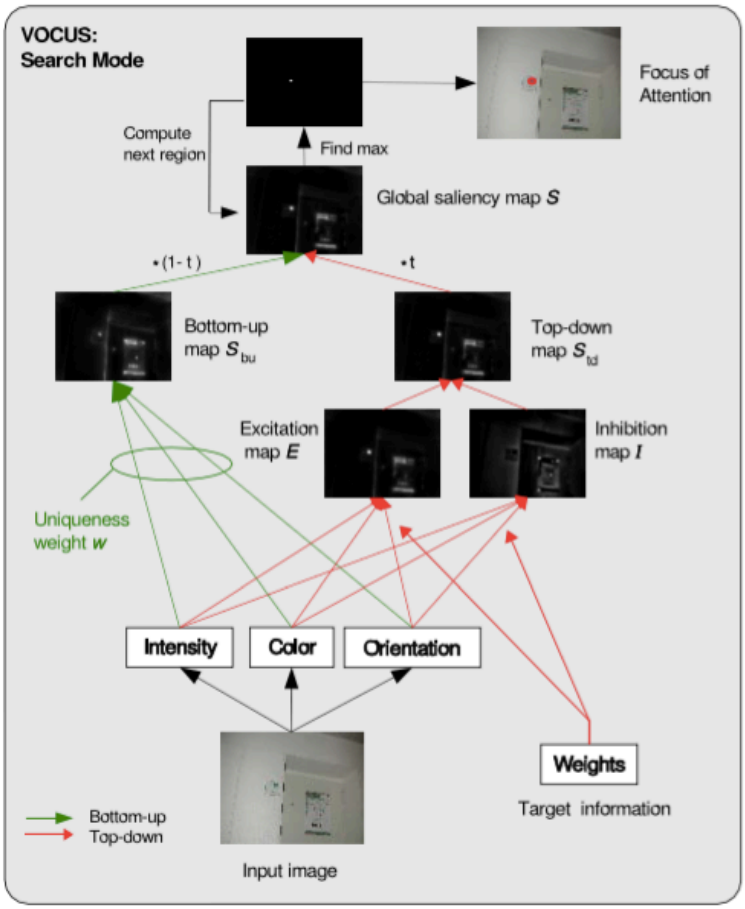

Figure 14. The searching mode of VOCUS (2006).

In searching mode, the model needs to find specified target. With the target's weight vector, feature maps of higher and lower importance are separated and summed into 2 maps (Exhibition and Inhibition maps) which are then later combined into a top-down map. Finally, global saliency map is obtained from weighted sum of bottom-up and top-down maps where the weight given to top-down map determines degree of concentration of the observer. Focus of attention is determined by the highest salience value in the global map.

3.4 Saliency Detection: A Spectral Residual Approach (2007) [8]

In this unique bottom-up model, proposed by Xiaodi Hou and Liqing Zhang, works in spectral domain. It uses the fact that an average of logarithm of spectrum representation is similar in many images, thus implies redundancies in visual information. By removing it, novel information about an input image is discovered as shown in Fig. 12.

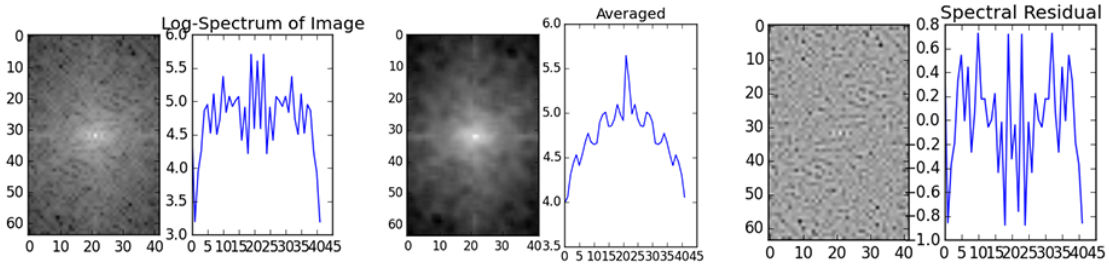

Figure 15. 1) Log-spectrum. 2) Average of log-spectrum. 3) Spectral residual.

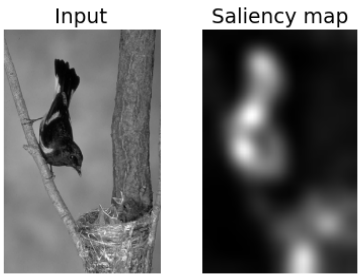

Figure 16. An input and a saliency map generated by Hou's model.

The process can be described briefly as the following steps:

- First convert image into spectral domain using Fourier Transform, and obtain only its amplitude.
- Take logarithm of the result in the first step, the result is called *log spectral representation* which can be plotted as curve of frequency and corresponding log intensity.
- Then apply uniform filter on the curve to obtain smoothed curve, which is assumed to be similarities shared with other images.
- Subtract the original curve with the smoothed curve, to emphasize only unique features of that object. This is called *spectral residual.*
- Finally convert the spectral residual back to spatial domain with Inverse Fourier Transform and then convolve with Gaussian kernel to get smoother values over regions. The result is saliency map.

An example of an input and the result is shown in Fig. 13. Note that this model is also available in OpenCV possibly due to its ease of implementation and its efficiency.

3.5 Frequency-tuned Salient Region Detection (2009) [9]

Developed by Achanta et al., the main idea is that saliency value at each location is the Euclidean distance between the color average of whole image in LAB color space and the color of image convolved with Gaussian filter. Euclidean distance between color vectors in LAB color space, unlike RGB color space, matches with perceptual difference in human vision. The process is illustrated in Fig. 14:

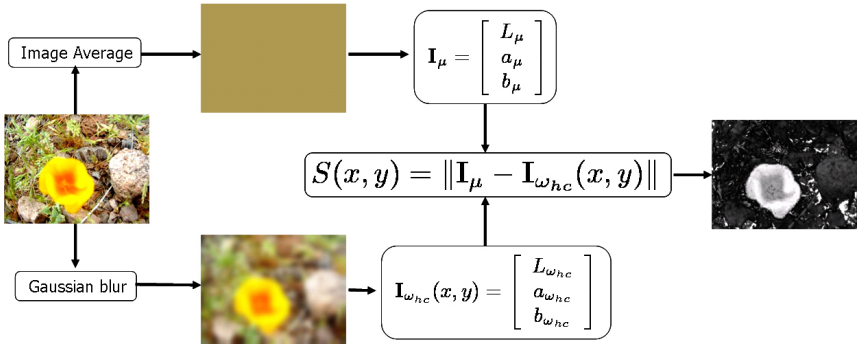

Figure 17. Achanta et al.'s model.

4. The Proposed Model

Based on the assumption that humans usually pay more attention near dense corner locations since they define contour of figures from ground, the core of our model is tentative corner extraction. The overview of the proposed model is illustrated in Fig. 18. First an input image is separated into intensity channel and 4 color opponency channels [5]. The intensity map is separated into 2 edge feature maps: on-center and off-center, to detect both bright and dark figures on its contrary backgrounds, according to [7]. For color channels, high-pass Laplacian filters are applied on each channel to obtain edge information. Laplacian filter is applied directly to image at original size, where edge corner information is most accurate. Next, these edge feature maps are convolved with Gabor filters at 4 different orientations, to express orientational selectivity in primary visual cortex (V1). Then, orientational responses on each channel are combined independently by multiplying these feature maps, to promote only locations with strong possibility for corner that has multiple orientational responses, and implies corner. The tentative corner maps are then suppressed by average of surrounding signals, normalized and finally summed to obtain a saliency map.

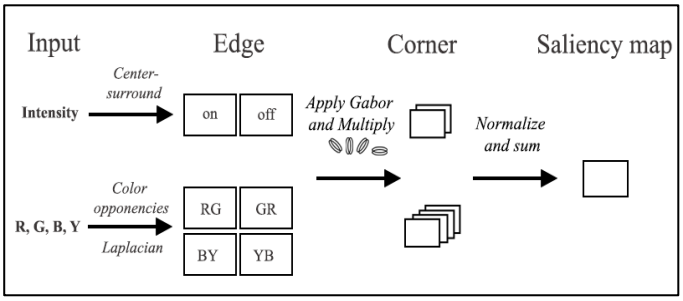

Figure 18. Diagram of the proposed model.

In the following sections: image representation, tentative corner extraction and channel combination, will describe each step in more details.

4.1 Image Representation

An input image is first separated into 4 channels: red (r), green (g), blue (b), and intensity channel (I) which is obtained by

$$I = \frac{r + g + b}{3}. \tag{1}$$

Intensity: On-center and Off-center are extracted using method similar to [7]. First Gaussian pyramids are constructed from intensity image, by downsampling a 5x5 Gaussian filter response of the image. In contrast to [7] where 5-band pyramid is used, our model use only first two bands. Each band is further divided into two intensity feature types: on-center-off-surround (or on-center) and off-center-on-surround (or off-center).

Center is determined directly from pixels at each location, and surround is computed by averaging all pixels in radius. On-center and off-center features given a scale (s) and a surround radius (rd) are obtained by:

$$Cs_{on,s,rd}(x,y) = \left\lfloor I_{s,rd}(x,y) - \frac{1}{N} \sum_{i,j \in [-rd,rd]} I_{s,rd}(x+i, y+j) \right\rfloor$$

$$Cs_{off,s,rd}(x,y) = \left\lfloor \frac{1}{N} \sum_{i,j \in [-rd,rd]} I_{s,rd}(x+i, y+j) - I_{s,rd}(x,y) \right\rfloor \quad (2)$$

Where I is the intensity image, s is level of Gaussian Pyramid, rd is radius of surrounding region, N is total number of pixels in surrounding area, and the operator $\lfloor \rfloor$ is half-wave rectifier which is defined as: $\lfloor x \rfloor = \max(0, x)$.

Since we assume baseline of any response starts at 0, negative values are ignored, in other words, neuron does not fire below this threshold.

This yields 4 maps for each intensity type (on and off), scale s = {1,2}, surround radius rd = {3,7}, so a total 8 maps are computed. For each type, maps at coarser scale are resized up to finer scale by linear interpolation and all are summed together pixel wise, yielding one on-center (I_{on}) and one off-center (I_{off}) map.

Color: In Lateral geniculate nucleus (LGN), color signals are also encoded by On- and Off-Center cells as intensity, but for two particular pairs of colors in opposite axis, which is called color opponency [5]. This is due to different wirings of three cone types in the retina each tuned for different wavelengths projected to LGN [10]. The opponency system consists of red-green (RG), green-red (GR), blue-yellow (BY) and yellow-blue (YB). For example, red-green denotes on for red at center, but off for green at surround. These channels are generated according to:

$$R = \left\lfloor r - \frac{g+b}{2} \right\rfloor$$
$$G = \left\lfloor g - \frac{r+b}{2} \right\rfloor$$
$$B = \left\lfloor b - \frac{r+g}{2} \right\rfloor$$
$$Y = \left\lfloor \frac{r+g}{2} - \frac{|r-g|}{2} - b \right\rfloor, \quad (4)$$

then yielding opponency maps are computed as follows:

$$RG = \lfloor R - G \rfloor$$
$$GR = \lfloor G - R \rfloor$$
$$BY = \lfloor B - Y \rfloor$$
$$YB = \lfloor Y - B \rfloor. \quad (5)$$

Laplacian filter is then applied independently to extract edge information for each color channel.

In total, there are 6 maps to represent edge information: two for intensity ($I_{on}$and I_{off}), and four for color opponencies (RG,GR,BY and YB).

An example of the process is shown in Fig. 19 where color channel R,G,B,Y are shown. The edge feature maps of channel RG, GR, BY, YB, $I_{on}$and I_{off} are shown in Fig. 20.

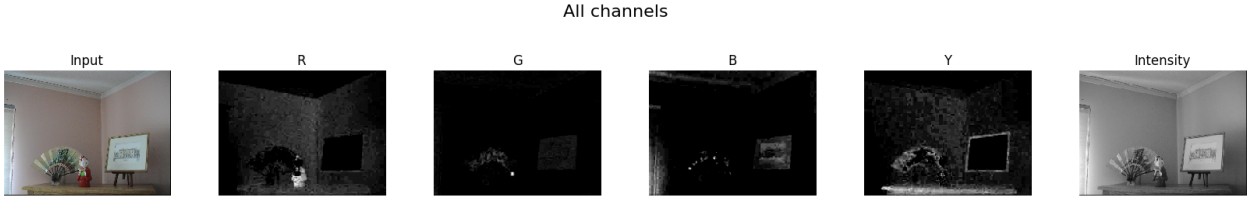

Figure 19. An input and its channels obtained by (4) for R, G, B, Y and (1) for intensity.

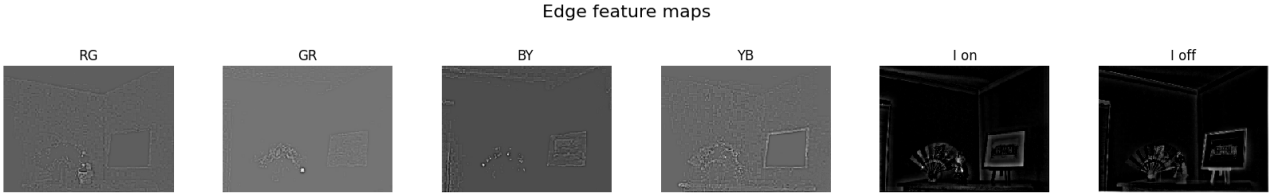

Figure 20. Six edge feature maps from the channels in Fig. 19. RG, GR, BY and YB are obtained from (5) followed with Laplacian filter. $I_{on}$and I_{off} are obtained from (2).

4.2 Corner Extraction

At this stage, tentative corner, line intersection and line ending locations are extracted by multiplying orientational feature maps together, as they usually respond to multiple orientations.

In V1, simple cells are tuned preferentially for different orientations that their receptive fields can be approximated with Gabor filter [11]. Each map from the previous step is convolved with Gabor filters at 4 orientations. Impulse responses of 2D Gabor filters are products of plane sinusoidal waves (cosine for real part and sine for imaginary part) and 2D Gaussian functions. Energy mechanism [12] is applied to yield an amplitude map (A) for each orientation:

$$A_{\Psi,\theta}(x,y) = \sqrt{G_{\Psi,c,\theta}(x,y)^2 + G_{\Psi,s,\theta}(x,y)^2}\,. \tag{6}$$

$G_{\Psi,c,\theta}$ and $G_{\Psi,s,\theta}$ are Gabor filter responses of input for real and imaginary parts respectively for input channel Ψ, and $\theta \in \{0°, 45°, 90°, 135°\}$. So four orientational feature maps are computed with the edge map from each channel. These four maps are then multiplied together directly to obtain tentative corner feature (Crn):

$$Crn_{\Psi}(x,y) = \prod_{\theta \in \{0°,45°,90°,135°\}} A_{\Psi,\theta}(x,y)\,. \tag{7}$$

However, this usually results in a highly sparse map for corner locations, since in orientational response in natural images is not as high as in artificial images. This is the reason why the Gabor filter cannot efficiently extract orientational information at cluttered scene where figures have no clear boundary. To promote more figural cues in a saliency map, natural logarithm function is applied on the response to obtain figural cue feature (Fc):

$$Fc_{\Psi}(x,y) = \ln(Crn_{\Psi}(x) + 1)\,. \tag{8}$$

It can also be viewed as adjusting the range of orientational responses with logarithm function and sum them up directly.

With this, six figural cue feature maps are obtained, one for each channel. Examples of Fc response for each channel are shown in Fig. 21.

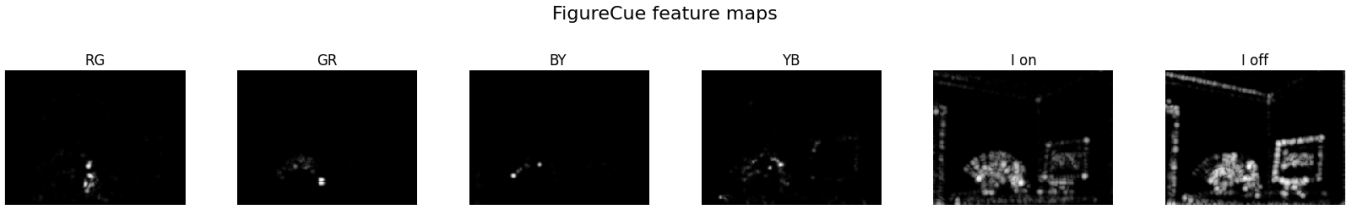

Figure 21. Figure cue feature maps obtained by (8) from the input in Fig. 19.

4.3 Channel Combination

Since figural cue responses for any channels are comparable, the maps are normalized into the same range of [0,1]. However, the figural cue responses in the same map are not equally important. Individual clusters of the response usually indicate main figure locations and are more salient. To promote individual cluster of figural cues, the response is subtracted from the average of surrounding pixels and clipped negative values with the half-wave rectifier:

$$Fc'_{\Psi}(x,y) = \left[Fc_{\Psi}(x,y) - \frac{1}{N} \sum_{i,j \in [-\rho w, \rho w]} Fc_{\Psi}(x+i, y+j) \right]. \quad (9)$$

N is number of pixels in the rectangular surrounding area with horizontal and vertical radius of ρw. Then the Fc' response map for each channel is adjusted its significance by dividing with the square root of number of local peaks as in [7] to bias the channel with few local maximums. Here ρw is one-tenth of the input image width and local peaks are calculated with minimum distance 3.

All six normalized figural cue maps of all corresponding channels: on and off intensity maps (I_{on} and I_{off}) and four color opponency maps (RG, GR, BY and YB) are combined together by pixelwise-addition, downsampled by linear interpolation and post-process by convolving with Gaussian filter. The saliency map generated is shown in Fig. 22 and other examples are in Fig. 23. All inputs and human fixation density maps are from the Toronto dataset [6].

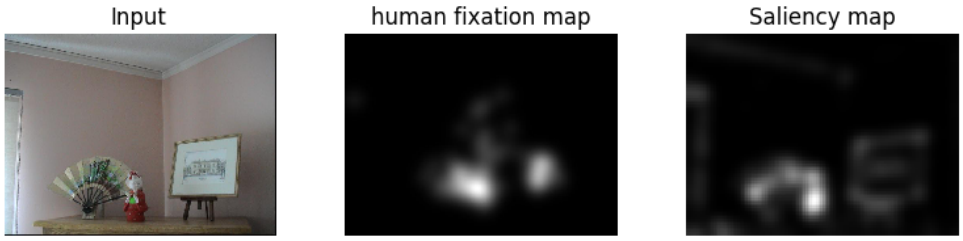

Figure 22. Input, the saliency map generated and a human fixation density map.

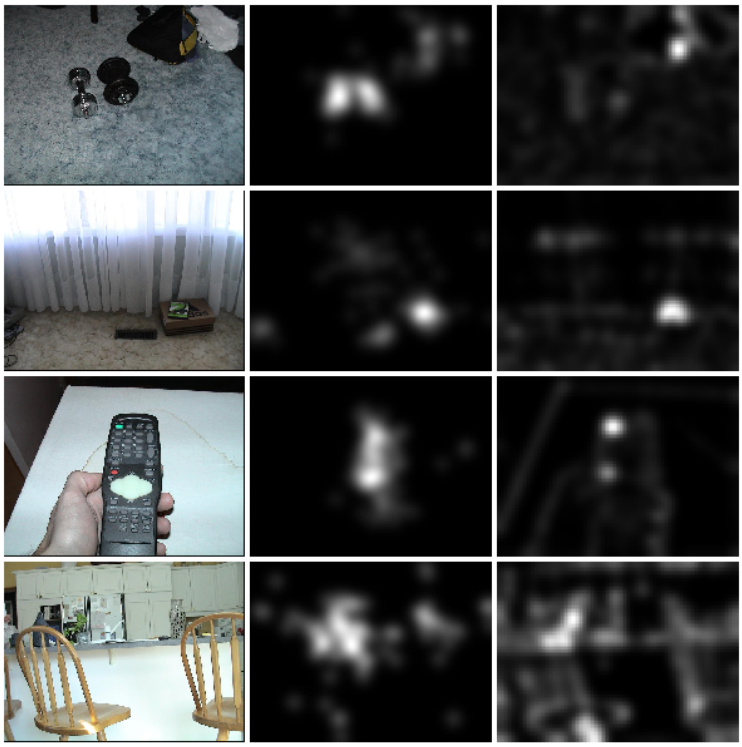

Figure 23. Other saliency maps generated by the proposed model. Columns from left to right: Input, human fixation density map and saliency maps. Inputs and human fixation maps are from the Toronto dataset.

5. Evaluation

The proposed model is tested on Toronto dataset [6] which contains 120 color images with resolution of 511 x 681, mostly indoor and outdoor environments. Random images are shown and eye fixations are recorded with viewing time 3 seconds on 20 subjects to obtain the ground truth for human fixation density.

5.1 Evaluation Metric

Two metrics used to evaluate the accuracy of the algorithm in literatures were surveyed.

5.1.1 Kullback-Leibler Divergence (KLD)

KLD is a measure of the difference between two probability distributions P and Q, by below formula:

$$D_{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(I)}$$

KLD has been used for saliency evaluation in [13], by computing difference between the histogram of saliency sampled at eye fixations and that sampled at random locations. However, it has been pointed out in [14] that this method is strongly sensitive to *center bias*.

Center bias refers to how viewer usually makes eye fixations at center of the image. So, the saliency map with only Gaussian blob at the center (Fig. 24) yields high KLD value regardless of image content.

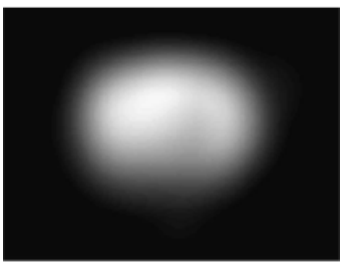

Figure 24. Gaussian blob: it has high KLD value regardless of input images.

This is because the histogram of saliency values at eye fixations (more probability to be in the center) is normally high, but that sampled randomly from the saliency map tends to be low, together it gives high KLD value.

In [14], Zhang et al., unlike the former method, uses KLD to compute the difference between the histogram of saliency sampled at eye fixations, and the one sampled at the same locations, but from **randomly selected** saliency map computed from **another image**. Most saliency values generated by the algorithm for an input image at real eye fixations from the dataset should be high, while these values at the same locations from another saliency map should be low as it is not the same image. By investigating the same locations (real eye fixations) between two histograms, it can avoid the default high difference value between fixations and random locations. Moreover by using saliency value from another saliency map, the metric can determine the meaningful relations between input image and saliency map produced with accuracy and detect the saliency map that performs by chance.

However, considering only probability density functions of two maps totally eliminates the correlation between spatial location information between them. Thus the alternative evaluation metric, area under ROC curve, is first introduced in [6].

5.1.2 The Original Area under the Receiver Operating Characteristic Curve (AUC)

Receiver Operating Characteristic curve (ROC curve) is a plot that illustrated performance of binary classifier. To evaluate a saliency map, the fixation map that contains fixation points on an image is collected. The saliency map determining the probability of fixations at each location is then computed.

To make a binary classification, a threshold is selected and applied on the saliency map to eliminate locations that have lower saliency value than the threshold. For example, 0.2 means only keep top 20 % salient value in the map, set value in these locations to 1 and the rest to 0. This process can be repeated with different threshold value, ranging from minimum (0) to maximum (1) value.

Next, four statistical definitions are introduced in context of saliency evaluation for a given thresholded saliency map:

True Positive (TP): correctly identified (a fixation and saliency value equals 1)
False Negative (FN): incorrect rejected (a fixation but saliency value equals 0)
False Positive (FP): incorrectly identified (not a fixation but saliency value equals 1)
True Negative (TN): correctly rejected (not a fixation and saliency value equals 0)

Then:

$$\textbf{True Positive Rate (TPR)} = \frac{TP}{TP+FN}$$

$$\textbf{False Positive Rate (FPR)} = \frac{FP}{FP+TN}$$

The sum of TP and FN is a positive set, or all human fixations recorded. The sum of FP and TN is a negative set, or other locations that are not fixations. Given a threshold applied on the saliency map and a fixation map, TPR and FPR can be determined. Repeat this process with threshold ranging from 0 to 1, and plot the FPR function of TPR to illustrate the performance of the classifier at all possible thresholds.

The area under the ROC curve (AUC or AUROC) can be computed as an area under the ROC curve to compare between different saliency algorithms. Higher value means better accuracy. AUC of chance level is 0.5 and AUC of perfect classifier is 1. An example of AUC computation is shown in Fig. 25.

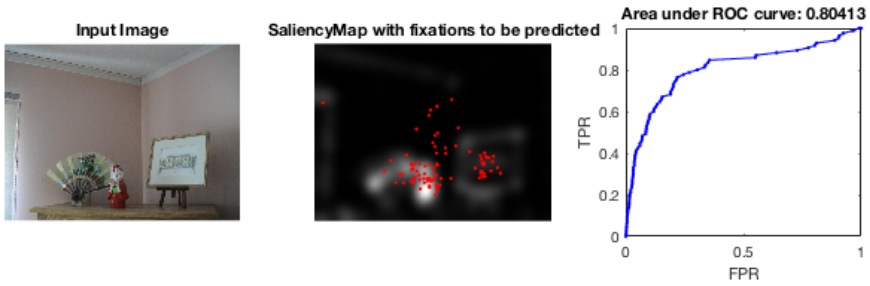

Figure 25. Input image and its evaluation with AUC. Red dots overlaying the saliency map in the second image denote human fixations from the dataset. AUC curve is shown in the third image.

The analysis and implementation of AUC in details can be further reviewed at [15].

However, AUC also suffers from center-bias as in KLD [14]. Gaussian blob at the center of image gives high AUC value and can be applied to increase performance of any algorithms significantly regardless of its output. It is illustrated in Fig. 26.

Figure 26. Gaussian blob always gives high AUC score.

First reason is center-bias, which human fixations usually gather around the center of an image and also due to the fact that positions of figures often placed at the center in the image frame. Another important reason is that although the positive set comes from human fixations, the negative set comes from the rest locations in the image, which is uniform distribution. So false positive rate can always be exploited by giving sparse saliency values around corner as in Gaussian blob.

5.1.3 Shuffled AUC

To workaround this bias, AUC-shuffled [14] instead collects eye fixations from other images (not containing fixations from the current image) and use them as negative set, in contrast with AUC that uses uniform distribution over all the image. Since possible human fixations are used, it implies that center bias is also in the negative set and helps adjust the false positive calculation.

By this, Gaussian blob on AUC-shuffled will give score around 0.5. An example of AUC-shuffled calculation is depicted in Fig. 27. Red dots (positive set) are human fixations of current input image and yellow dots (negative set) are human fixations selected randomly from other maps. It can be seen that the negative set also affected by center-bias as the positive set. Thus, AUC-shuffled provides more fair calculation of both true positive rate and false positive rate. It will be used as a metric to evaluate the proposed model in the next section.

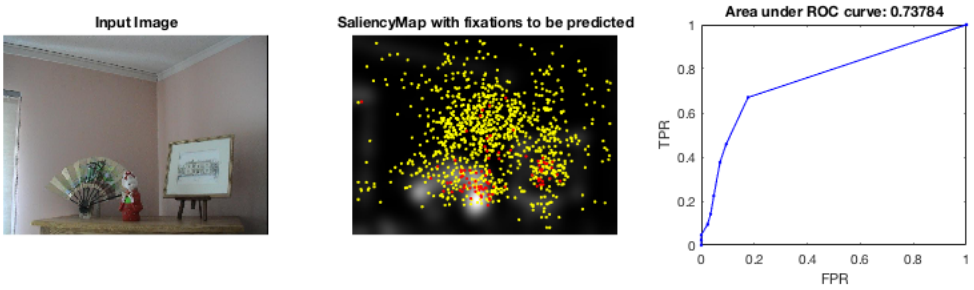

Figure 27. AUC-shuffled calculation.

5.2 Experimental Results

The proposed model is tested on Toronto dataset [6] which contains 120 color images with resolution of 511 x 681, mostly indoor and outdoor environments. Random images are shown and eye fixations are recorded with viewing time 3 seconds on 20 subjects to obtain the ground truth for human fixation density.

Although AUC-shuffled can evaluate performance from spatial information and is not suffered from center-bias, AUC-shuffled is still sensitive to final blurring on a saliency map [16]. To evaluate the overall performance of a model, final saliency maps are convolved with Gaussian kernel at

various standard deviations (STD). Average of AUC-shuffled scores of 120 images is then plotted as function of standard deviations of the Gaussian kernel.

Other six saliency models (Itti [5], GBVS [17], AIM [6], SUN [14], SUN-small, Signature [16]) are compared with our model denoted as Cor, using publicly available codes at authors' website. Note that SUN-small used input images in smaller size than SUN (one-tenth of original image size). The relationship of standard deviations of final blurring to average of AUC-shuffled scores on 120 images is shown in Fig. 28. Mean and optimal STD scores for each model are also shown in Table I.

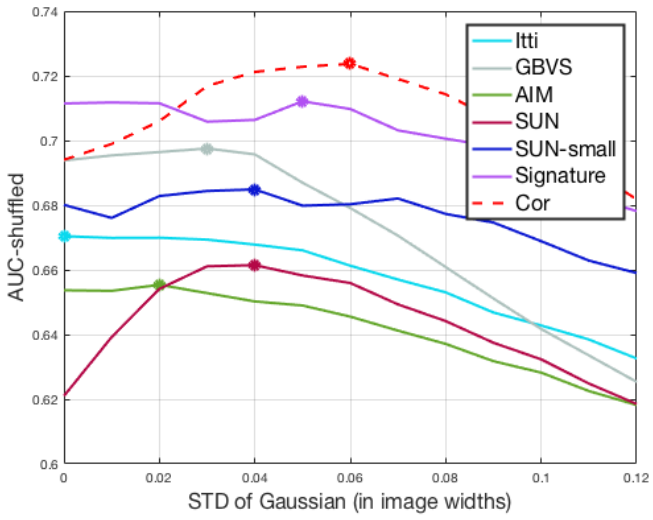

Figure 28. The plot of average of AUC-shuffled scores of 120 images on Toronto dataset as a function of standard deviations of Gaussian kernel (in image widths) used in final blurring of saliency map generated. The dots represent optimal STD values for each model.

Table 1. Performance of 7 algorithms on Toronto dataset.

Algorithm	**AUC-shuffled (mean STD)**	**AUC-shuffled (optimal STD)**
Itti et al. [5]	0.6573	0.6704
GBVS [17]	0.6714	0.6975
AIM [6]	0.6415	0.6554
SUN [14]	0.6429	0.6615
SUN-small	0.6764	0.6849
Signature [16]	0.7016	0.7121
Cor	**0.7074**	**0.7237**

We have mentioned in the proposed model section that we do not use Crn feature (7) directly in our model because corner feature map generated is very sparse. Moreover corner responses are clearly deviated from figure areas where people usually look at. Corners define contours of figures and are glanced only for a short time to further extract figures.

Our method (denoted as Cor) used (8) for figural cue. Adjusting responses by logarithm allow more corners to be seen. Gaussian filtering at final stage also helps to cover areas adjacent to corner response that implies figure locations, as shown in Fig. 28. Our model significantly increased in performance initially as more STD of Gaussian blur is used and reached the peak at STD of 0.06 of the image width.

Overall, our saliency model based on figural cue can suggest tentative figure locations and the model is very competitive among other approaches.

6. Conclusion

In this chapter the implications, strength, weakness and motivation behind the proposed model is discussed in section 6.1. The chapter ends with section 6.2, which describes possible further improvements on the proposed model.

6.1 Discussion

Firstly, in the proposed model, we generate saliency map from tentative corner features. It must be emphasized that our simple corner feature extraction method is neither sufficiently accurate to be used in natural images since it gives highly sparse response for corners, which can be detected only in artificial settings, nor good enough to be used in place of other corner feature extraction algorithms. But it is good for our work because we use logarithm function on each pixel to adjust the response and show more corners. In low resolution, however, locations of these features often reside in main figures in the natural scene, and can be used to approximate the attention of the viewer.

Secondly, saliency models based on Feature Integration Theory ([5], [7]) compute intensity, colors and orientation features in parallel, and finally distinct feature maps are promoted by apply weighting and normalization functions. In contrast, our model does not compute any conspicuity directly on each channel in parallel and thus it does not exhibit similar results to psychological tests such as orientational “pop out” as in other models since our model operate mainly on corner features to probe for possible figure locations.

Last but not least, the main reason we consider corner and line intersection as features for saliency modeling is because of its possibility to generalize and be used in higher-order visual processing. Apart from intensity, colors and orientation features extracted from the retina, LGN and up to V1, the use of corner, line intersection and line ending plays an important role in intermediate shape processing in V4 [1]. By encoding corners after orientations together with further interactions of neurons such as long-range horizontal excitation mechanism in V1 [18], some of human visual phenomenon like illusory contours can be achieved.

This phenomenon is emphasized in Fig. 29 and 30, corner features are extracted from input images into corner maps using (7). Each channel is normalized into range [0,1] and sum up directly.

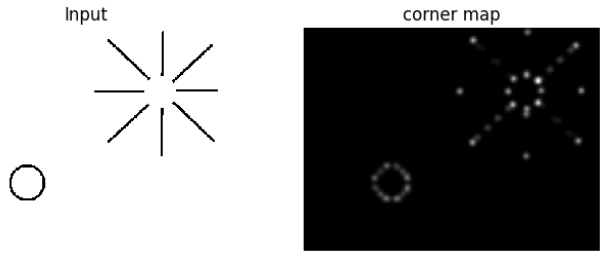

Figure 29. Corner maps that can be processed further to perceive the upper-right figure as a circular shape with illusory contour. The lower-left circle is there for comparison.

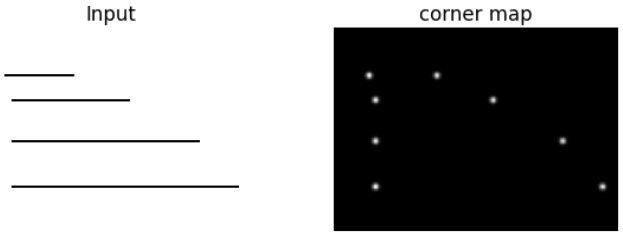

Figure 30. Corner maps for lines at different lengths. The corner features can imply the length of a line and remove redundancy at intermediate intervals.

Fig. 29 shows that corner and line ending feature responses are generalized to cover both circular shape on the lower left, and implicit circular shape on the upper right that can be detected later on with illusory contour mechanism. In this way, connecting dots along contour can help uncover illusory shape perceived by humans. However, prior knowledge about shapes is still needed to connect these dots correctly.

Similarly, a type of shape might be line. Fig. 30 shows corner responses for line of variable lengths. The intermediate visual information along the lines may not be equally important as a pair of corner responses at the beginning and the end, since they directly determine the length of the line. Note that this was not meant to ignore those visual responses at intermediates completely since endpoints alone cannot sufficiently determine shape. Both orientation and corner information must be integrated to yield accurate shape representation for further visual processing.

6.2 Further Improvements

We proposed a bottom-up saliency model based on corner features to suggest possible figure locations in a natural image. Further processing can improve the model either by using better corner extraction algorithm to efficiently extract corners in natural scenes, or employing a more sophisticated method to predict figure locations near those corners. Moreover, traditional biological saliency models operate on three common low-level features: intensity, colors and orientation channels. Middle-level feature like corner, line endings or curves might be the fourth important link to connect saliency computation from low-level to higher-level features like shape.

7. References

[1] A. Pasupathy and C. E. Connor, “Responses to contour features in macaque area V4,” *J. Neurophysiol.*, vol. 82, no. 5, pp. 2490–2502, Nov. 1999.

[2] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[3] C. Koch and S. Ullman, “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry,” in *Matters of Intelligence*, L. M. Vaina, Ed. Springer Netherlands, 1987, pp. 115–141.

[4] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, “A model of proto-object based saliency,” *Vision Res.*, vol. 94, pp. 1–15, Jan. 2014.

[5] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 11, pp. 1254–1259, 1998.

[6] J. K. T. Neil D. B. Bruce, “Saliency Based on Information Maximization.,” *Adv Neural Inf Process Syst*, vol. 18, 2005.

[7] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, vol. 3899. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[8] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*, 2007, pp. 1–8.

[9] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, 2009, pp. 1597–1604.

[10] L. M. Hurvich and D. Jameson, “An opponent-process theory of color vision,” *Psychol. Rev.*, vol. 64, Part 1, no. 6, pp. 384–404, Nov. 1957.

[11] “Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, 1233-1258.” [Online]. Available: https://www.researchgate.net/publication/19719670_Jones_J_P_Palmer_L_A_An_evaluation_of_the_two-dimensional_Gabor_filter_model_of_simple_receptive_fields_in_cat_striate_cortex_J_Neurophysiol_58_1233-1258. [Accessed: 20-Mar-2016].

[12] E. H. Adelson and J. R. Bergen, “Spatiotemporal energy models for the perception of motion,” *J. Opt. Soc. Am. A*, vol. 2, no. 2, pp. 284–299, Feb. 1985.

[13] L. Itti and P. F. Baldi, “Bayesian Surprise Attracts Human Attention,” in *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*, Cambridge, MA, 2006, pp. 547–554.

[14] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics,” *J. Vis.*, vol. 8, no. 7, pp. 32.1–20, 2008.

[15] O. Le Meur and T. Baccino, “Methods for comparing scanpaths and saliency maps: strengths and weaknesses,” *Behav. Res. Methods*, vol. 45, no. 1, pp. 251–266, Mar. 2013.

[16] X. Hou, J. Harel, and C. Koch, “Image Signature: Highlighting Sparse Salient Regions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.

[17] “Graph-Based Visual Saliency - 3095-graph-based-visual-saliency.pdf.” [Online]. Available: https://papers.nips.cc/paper/3095-graph-based-visual-saliency.pdf. [Accessed: 20-Mar-2016].

[18] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, “Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex,” *J. Neurosci.*, vol. 17, no. 6, pp. 2112–2127, Mar. 1997.