國立交通大學

電子工程學系 電子研究所碩士班

碩士論文

基於影像特徵對之顯著區域偵測技術

Salient Region Detection Based on Image Feature-Pair Distributions

研 究 生：黃文中

指導教授：王聖智 博士

中華民國九十八年七月

基於影像特徵對之顯著區域偵測技術

Salient Region Detection Based on Image Feature-Pair Distributions

研 究 生：黃文中　　Student：Wen-Chung Huang

指導教授：王聖智博士　　Advisor：Dr. Sheng-Jyh Wang

國立交通大學

電子工程學系 電子研究所碩士班

碩士論文

A Thesis

Submitted to Department of Electronics Engineering & Institute of Electronics

College of Electrical and Computer Engineering

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of Master

in

Electronics Engineering
July 2009
Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

基於影像特徵對之顯著區域偵測技術

研究生：黃文中　　　指導教授：王聖智 博士

國立交通大學

電子工程學系　電子研究所碩士班

摘要

在本論文中，我們提出一套偵測人眼視覺顯著區域之技術。給定一張影像，藉由我們提出的演算法可以立即判斷出哪些位置區塊會是人眼較易去注意的地方。輸入影像會先被拆解成三種通道，包含了強度跟兩個對比色彩通道。對於個別通道，會將其建構成特徵對的分布圖，並藉由分析特徵對分布圖的結果反映射回空間域去識別出視覺顯著區域。此外，為了抑制雜訊造成的影響，我們另外加上了正規化的步驟，以提高顯著區域劃分的成功率。根據實驗結果，我們發現此技術確實可以偵測出人眼視覺的顯著區，同時過濾掉較不重要的資訊。

Salient Region Detection Based on Image Feature-Pair Distributions

Student: Wen-Chung Huang Advisor: Dr. Sheng-Jyh Wang

Department of Electronics Engineering, Institute of Electronics

National Chiao Tung University

Abstract

In this thesis, we propose an algorithm for the detection of human visual saliency regions. Given an image, the proposed algorithm can automatically determine these locations where humans tend to pay more attention to. The image is first decomposed into three channels, including one intensity channel and two opponent-color channels. For each channel, a feature-pair distribution is created for saliency analysis, and the analysis result is mapped back to the spatial domain to identify visually salient regions. Beside the suppression of noise interference, a normalization stage is included to improve the performance of detection. As demonstrated in the experimental results, the proposed method can successfully identify visual saliency regions in human visual reception and, at the same time, filter out less crucial information.

誌謝

這兩年的研究所生涯，我得到的不只是專業知識的養成，更學會了如何去面對問題、了解問題，進而解決問題。在此要特別感謝我的指導教授 王聖智老師，除了對學術的堅持，以及對我們的信任，讓我們可以在自由的研究環境下發揮，並不忘細心的教導我們許多待人處事的道理；同時要感謝 辛正和老師，每次的會議總是很有耐心的教導並且給予我許多建議。此外，很感謝交大傳播科學系的 陶振超老師租借我們使用眼動儀儀器，並提供了在我專業之外的建議及協助，也很感謝婉雲、孟琪、佩瑩及傳科所的大家幫忙我完成最後的眼動人因實驗。也感謝實驗室的學長姐、同學以及學弟們，不管遇到什麼事情大家都是很熱心的幫助解決或是一同面對。感謝敬群學長、慈澄學長和禎宇學長總是適時的給予建議，讓我突破許多研究上的瓶頸；感謝瑞男、庭瑋和維辰在這條路上總是給我諸多幫助，不管是研究上或是生活上。同時也要謝謝我的家人，你們永遠是我最強力的後盾，讓我能無後顧之憂的勇往直前；最後要感謝我的女朋友，文婷，因為這一路上有妳的陪伴跟支持，讓我在面對挫折時更加有勇氣去面對、克服，也因為妳的鼓勵，成為了我一路向前的動力來源。還有許許多多的朋友們，沒有你們大家，就沒有這本論文的完成，真的是萬分感激！謝謝你們！

Content

List of Figures

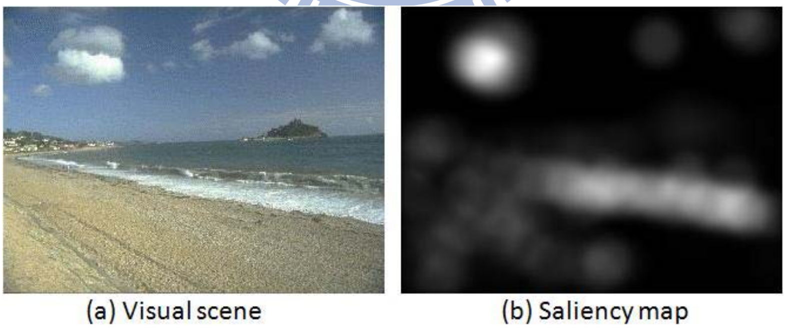
List of Tables

Chapter 1.

INTRODUCTION

The first step towards object recognition is object detection. Object detection aims at extracting objects from the background before recognition. However, before performing recognitive feature analysis, how can a machine vision system extract the salient regions from an unknown background?

Due to the complicated processing in human visual perception, humans don't process the whole visual field as a scene come to their eyes [1]. Instead, humans tend to selectively focus on certain targets that they are more interested in. To mimic this mechanism, a few researchers had suggested the use of different visual features for the formation of a topographically oriented map, the so-called saliency map. An example of the saliency map is shown in Figure 1-1, where brighter areas indicate visually more salient regions. These visually salient areas are generally regarded as the candidates of visual attention in human eyes. The detection results of visual saliency map can thus provide useful information for efficient detection of interested targets in a complicated scene.



(a) Visual scene

(b) Saliency map

Figure 1-1 *A visual scene and its corresponding saliency map*

In computer vision, many models have been proposed to simulate the behavior of eyes, such as SaliencyToolBox (STB), Neuromorphic Vision Toolkit (NVT), and etc.. However, these methods demand high computational cost and their remarkable results mostly rely on a proper choice of parameters. In 2007, a simple and fast approach based on Fourier transform, called Spectral Residual (SR), was proposed. This method used SR of the amplitude spectrum to obtain the saliency map. In 2008, another method had been proposed which used Quaternion Fourier Transform to deal with phase spectrum (QPFT). Their method performed well at detecting salient objects. However, if the size of the processing image is not properly chosen, the resulting output saliency map may just become the result of edge detection.

In this thesis, we propose a simple and efficient algorithm for saliency region detection. First, an input image is decomposed into three different channels: intensity, RG color, and BY color. A feature-pair distribution is used to analyze the semi-global information of each channel. With the feature-pair distributions, the saliency weight of each pixel in the original image is estimated to form the conspicuity maps. After all three conspicuity maps are obtained, the normalization step is taken to suppress noise interference and irrelative regions. The normalized conspicuity maps are then merged in a data-driven manner to form the final saliency map. The proposed saliency region detector does not require complex computations. As will be shown in the experimental results, the proposed system can be applied to various kinds of images to obtain visual saliency regions that are consistent with subjective observations.

This thesis is organized as follows. In Chapter2, we introduce the background of existing saliency map models and related techniques. In Chapter3, we present the proposed image feature pair-distribution method. Experimental results are shown in Chapter4. Finally, we will make a brief conclusion in Chapter5.

Chapter 2.

BACKGROUNDS

In this chapter, we will introduce a few salient region detecting approaches developed in recent years. Firstly, a brief introduction to the definition of saliency map is presented in Section 2.1. Next, related models for the creation of saliency map, together with the functional taxonomy, will be introduced in Section 2.2. Since our method is based on Itti’s visual saliency model and image feature-pair distributions, related concepts and their origins are mentioned in Section 2.3 and 2.4, respectively.

2.1. VISUAL SALIENCY MAP

The saliency map is a topographically arranged map that represents visual saliency of a corresponding visual scene. One of the most severe problems of perception is information overload. Peripheral sensors generate afferent signals more or less continuously and it would be computationally costly to process all the incoming information all the time. Thus, it is important for the neural system to make decisions on which part of the available information to be selected for subsequent processing, and which part to be discarded. Furthermore, the selected stimuli need to be prioritized, with the most relevant being processed first while the less important ones later. This selection and ordering process is called selective attention. Among many other functions, attention to a stimulus has been considered necessary for conscious perception.

What determines which stimuli to be selected by the attentional process and which to be discarded? Many interacting factors contribute to this decision. It has proven useful to distinguish between bottom-up and top-down factors. The former are those that depend only on the instantaneous sensory input, without taking into account the internal state of the organism. Top-down stimuli, on the other hand, does take into account the internal state, such as the goals the organisms has at this time, personal history and experiences, etc. A dramatic example of a stimulus that attracts attention using bottom-up mechanisms is the case of a fire-cracker going off suddenly. An example of top-down attention is the focusing onto some difficult-to-find food items by a hungry animal, which may ignore most "salient" stimuli but food.

2.2. MODELS OF VISUAL SALIENCY MAP

In general, the existing visual saliency detection algorithms can be classified into bottom-up approaches and top-down approaches, depending on whether the prior knowledge of the visually attended objects is used. A top-down approach usually requires some prior knowledge of the targets in order to extract task-dependent clues. However, in practice, the prior knowledge of the targets objects is usually unavailable. Hence, in this paper, we mainly focus on the development of a bottom-up approach.

According to the survey in [2], a bottom-up approach typically consists of the following functional modules:

- Extraction

  Feature vectors, which may include intensity, color double-opponent, orientation, etc, are extracted at different locations of the image plane.
- Activation

  Based on the extracted feature vectors, a few conspicuity maps are formed to identify the candidates of visual saliency regions.
- Normalization

  Each conspicuity map is normalized to emphasize its prominent regions.
- Combination

  All conspicuity maps are combined into the final visual saliency map.

In this thesis, we propose a simple and efficient algorithm for saliency region detection. In this system, we focus mainly on the activation and the combination modules. The module of activation is designed to obtain more reliable feature information for saliency region detection. Due to severe signal-to-noise problem, combination of several feature maps may lead to wrong salient region detection. In the following discussion, we will focus on a few algorithms which are related to feature combination.

2.2.1. Bottom-Up Models

Without training, human vision can focus on general salient objects rapidly in a clustered visual scene because of the existence of visual attention mechanism. The study of this visual attention mechanism has become an intriguing subject for more and more researches.

In the past decade, several computational models have been proposed to simulate human's visual attention model. Koch and Ullman presented in [3] a popular computational model for visual attention mechanism. Their biologically plausible model is purely data-driven and requires only image data. In their approach, four major principles are adopted: visual attention acts on a multi-featured input; saliency of locations is influenced by the surrounding context; the saliency of locations is represented on a scalar map; and both the winner-take-all and inhibition of return are suitable mechanisms for attention shift, as illustrated in Figure 2-1.

On the other hand, Itti *et al*. proposed in [4] a much more complicated system, which decomposes an input image into a set of distinct channels, such as luminance, different colors, and different orientations. They further used Gaussian pyramids to obtain feature maps of different scales. Each feature is then computed in a center-surround manner akin to human's visual receptive fields. For each feature type, the multi-scale feature maps are combined in a competitive way to form a unique conspicuity map. All the conspicuity maps are then integrated into a single saliency map, over which the winner-take-all rule and the inhibition-of-return mechanism are applied.

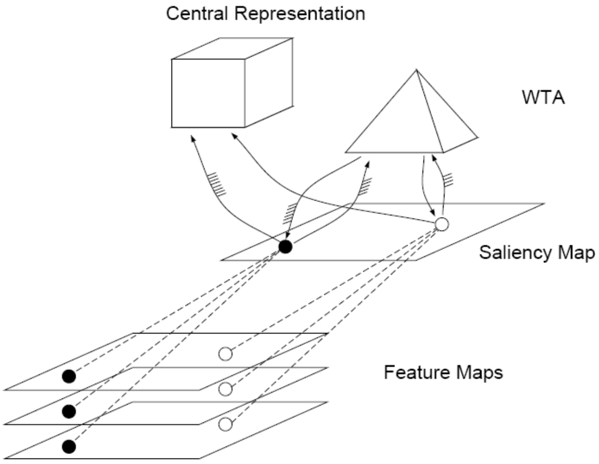


***Figure 2-1** The saliency-based model of visual attention as suggested by Koch and Ullman [3].*

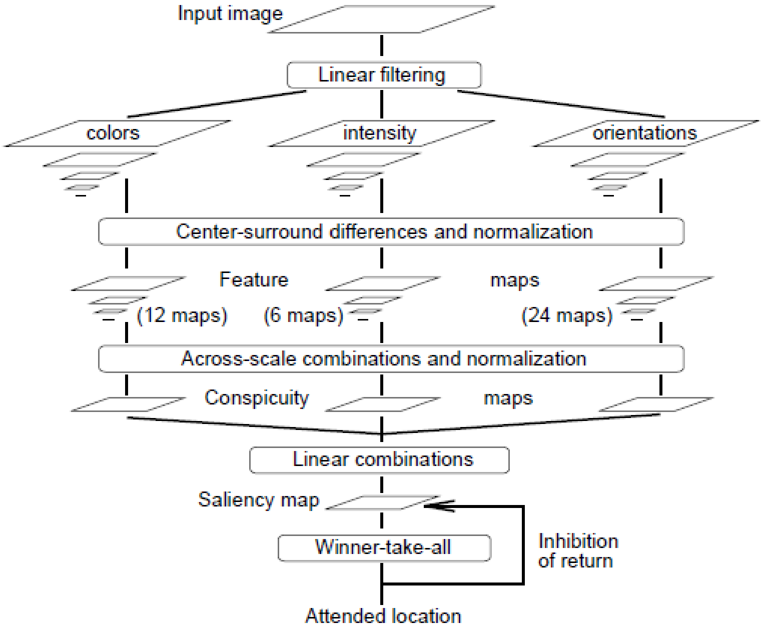


Figure 2-2 *General architecture of the model proposed by [4].*

Following Rensink's theory [5], Walther further extended Itti's model to handle "proto object" and built the SaliencyToolBox (STB) in [6]. In this extended model, Walther *et al*. proposed a feedback loop to automatically form proto objects and to search for proto objects in natural scenes, as shown in Figure 2-3. Even though both Itti's and Walther's models performed quite well in detecting the visual saliency regions in some images, these methods demand high computational cost and their remarkable results usually rely on a proper choice of the controlling parameters.

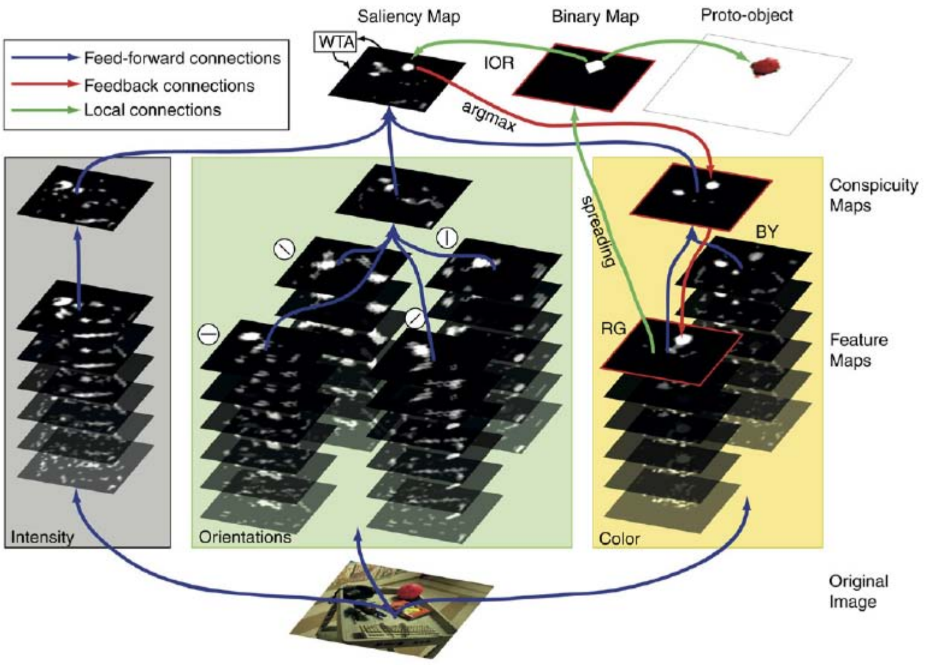
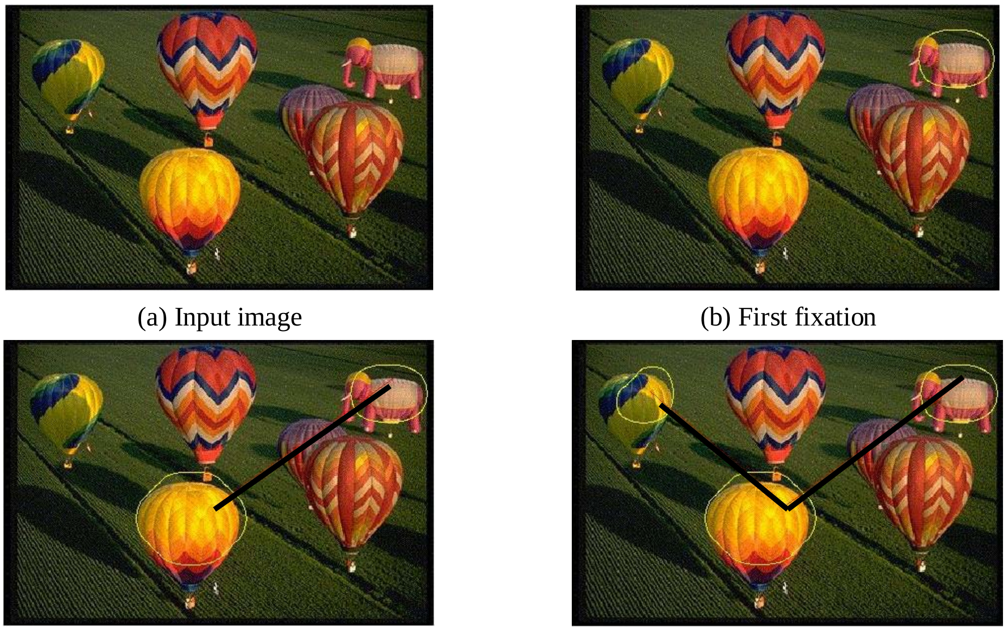


Figure 2-3 *Illustration of the processing steps for obtaining the attended region.[6]*



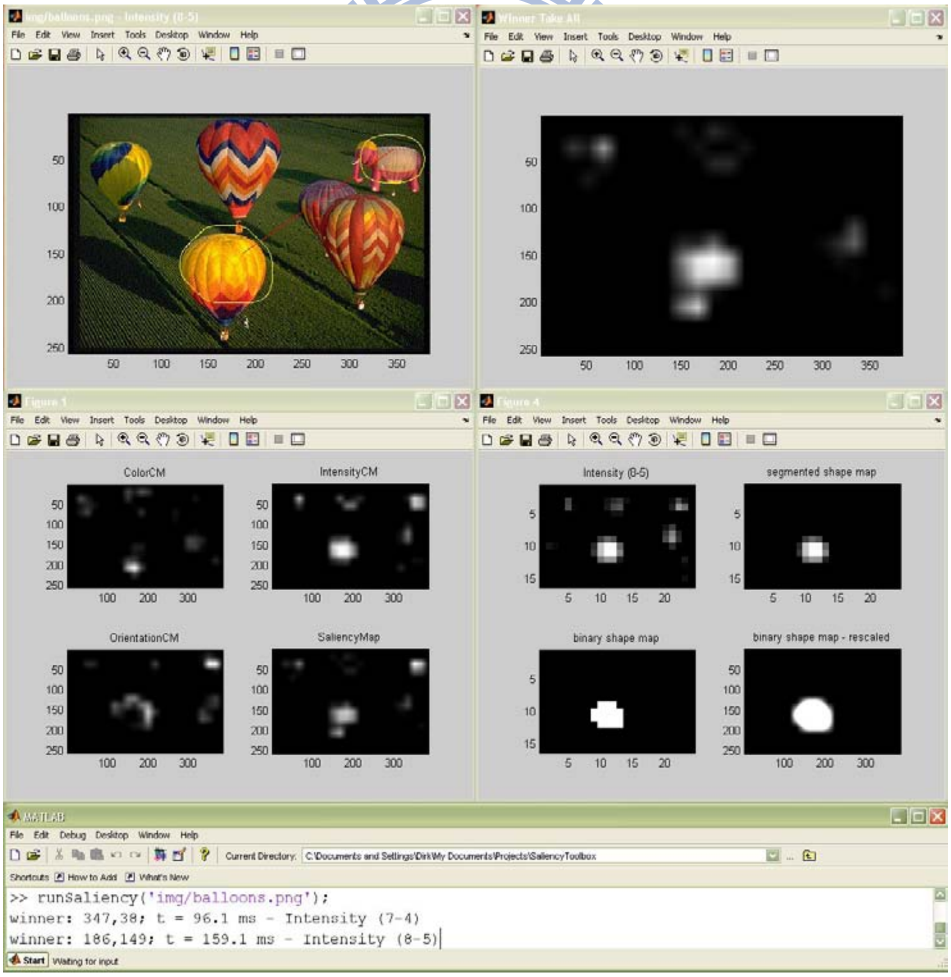
(a) Input image

(b) First fixation

(c) Second fixation

(d) Third fixation

Figure 2-4 *Result of Walther's proposed model with shift attention.*



img/balloons.png - Intensity (8-5)
Winner Take All
Figure 1
ColorCM
IntensityCM
OrientationCM
SaliencyMap
Figure 4
Intensity (8-5)
segmented shape map
binary shape map
binary shape map - rescaled
MATLAB
File Edit Debug Desktop Window Help
Current Directory: C:\Documents and Settings\Dirk\My Documents\Projects\SaliencyToolbox
Shortcuts How to Add What's New
Start Waiting for input

```
>> runSaliency('img/balloons.png');
winner: 347,38; t = 96.1 ms - Intensity (7-4)
winner: 186,149; t = 159.1 ms - Intensity (8-5)
```

Figure 2-5 *Screen shot of a typical display while running the SaliencyToolbox.[6]*

Recently, in 2007, a simple and fast approach based on Fourier transform, was proposed in [7]. This method is named Spectral Residual (SR) and is based on the simple Fourier Transform operation. In this approach, the authors calculated the residual of the log amplitude spectrum of the given image with respect to a reference profile to obtain the spectral residual, as shown in Figure 2-6. The saliency map is then obtained by transforming the spectral residual back to the spatial domain. In spite of its simple operation, this SR method performed surprisingly well in detecting the saliency regions of many images. All these models mentioned above, however, only consider static images.

However, after careful analysis, Guo pointed out in [8] that the spectral residual of the log amplitude spectrum is actually not essential to the calculation of the saliency map. Instead, the phase spectrum plays the major role in detecting saliency regions, as compared in Figure 2-8. In Guo's approach, each pixel of the given image is represented by a quaternion that consists of color, intensity, and motion features. The phase spectrum of the Quaternion Fourier Transform (QFT) is calculated and is used to obtain a spatio-temporal saliency map. Even though these spectrum-based approaches can detect saliency regions in a very efficient way, their detection results are actually more like the results of boundary detection.

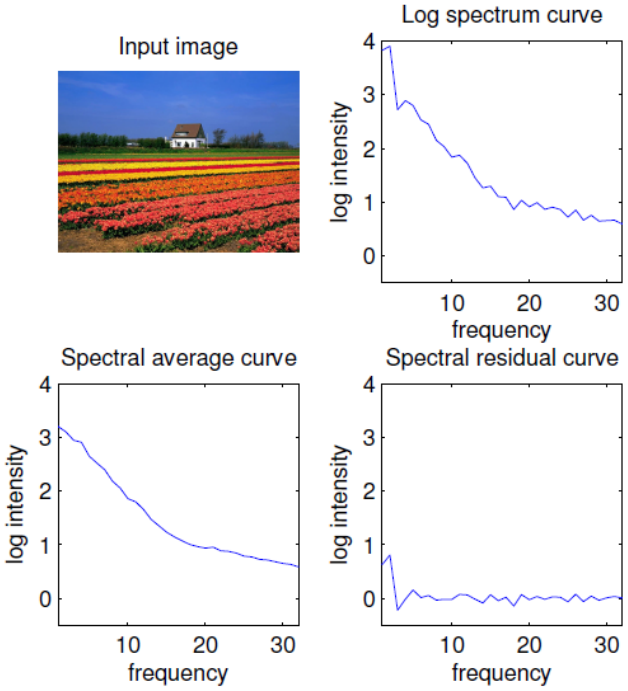


Figure 2-6 *The difference (SR) between the original signal and a smooth one in the log amplitude spectrum.[7]*

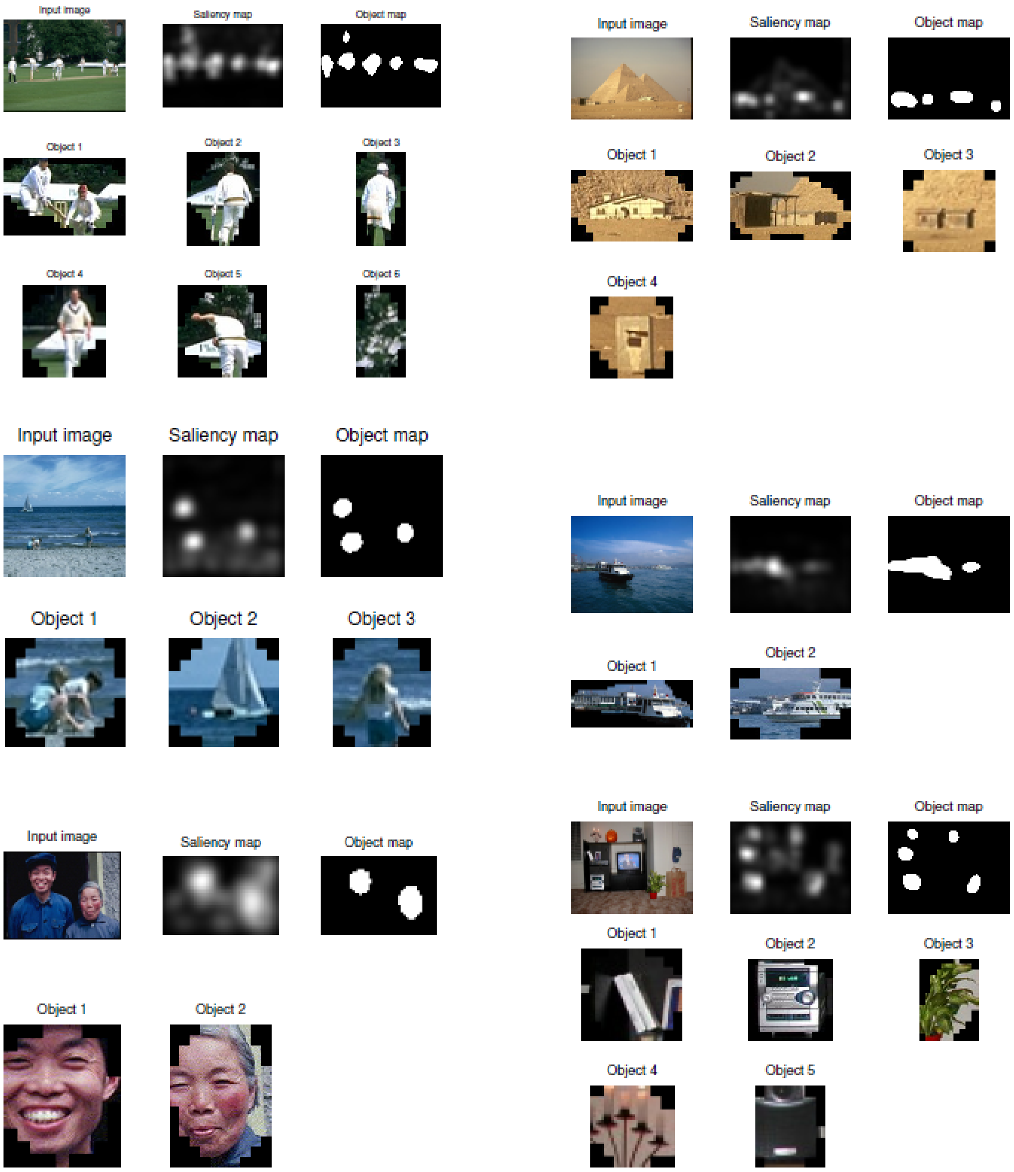


Figure 2-7 *Detecting objects from input images.*

Objects are popped up sequentially according to their saliency map intensity.[7]

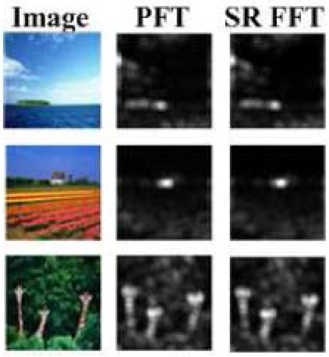


Figure 2-8 *Test results from three input images.*

(left) Input images, (middle) Saliency maps from PFT, (right) Saliency maps from SR[8].

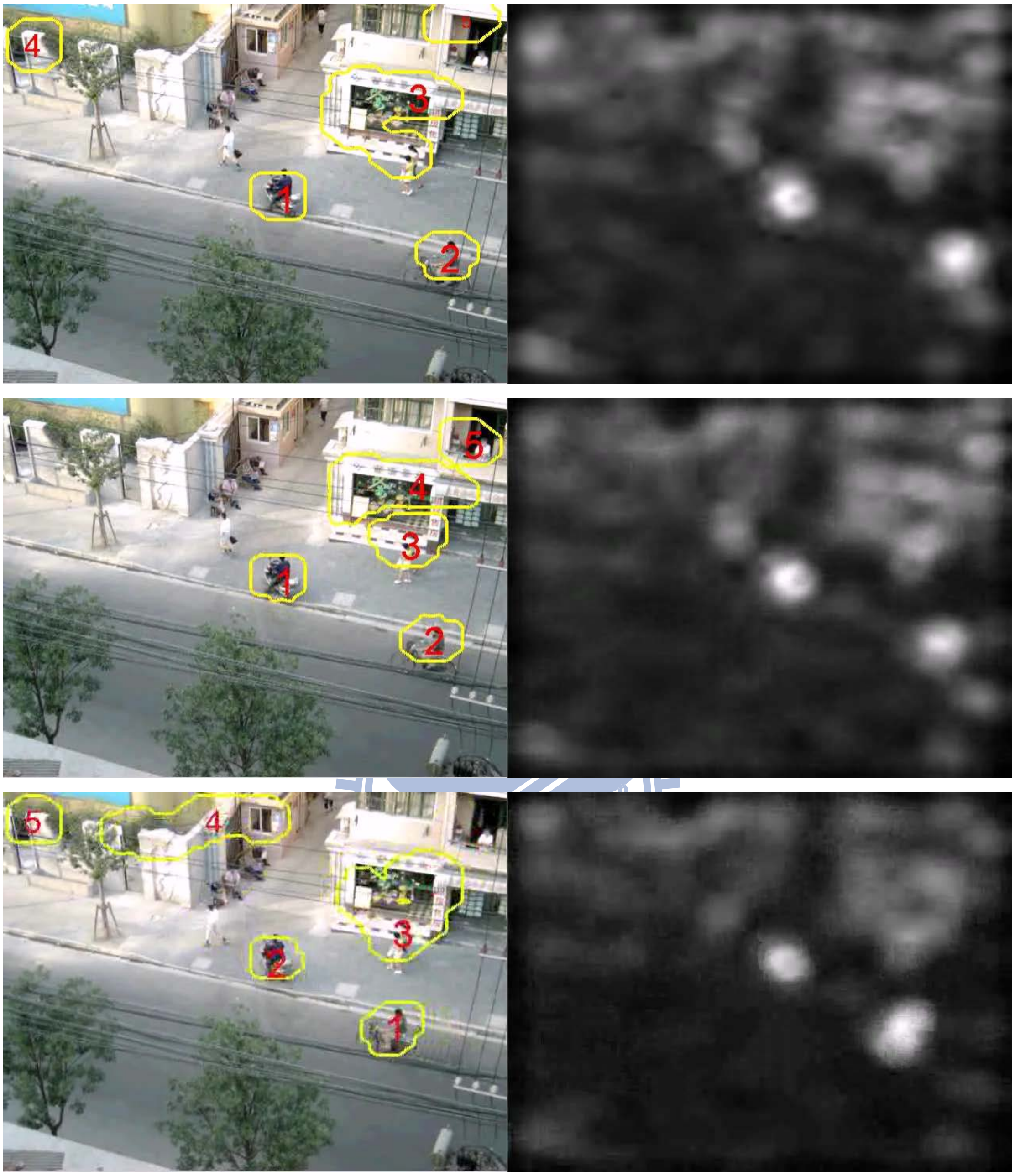


Figure 2-9 *Resulting sequence of PQFT as proposed by [8]*

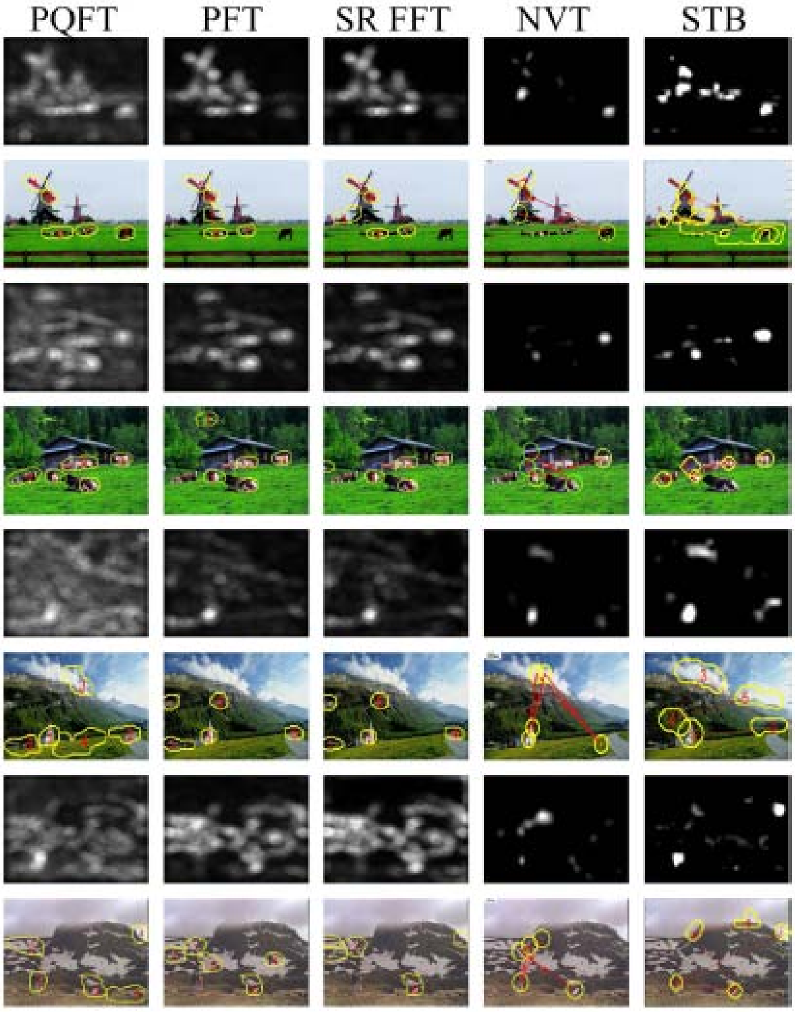


Figure 2-10 *Comparison of five models in four natural images [8].*

2.3. THE SALIENCY-BASED MODEL OF VISUAL ATTENTION

As already mentioned in Section 2.2.1, the saliency-based model of visual attention has been presented by Koch and Ullman in [3]. This model is based on four major principles: visual attention acts on a multi-featured input; saliency of locations is influenced by the surrounding context; the saliency of locations is represented on a scalar map -- the saliency map; and the winner-take-all and inhibition of return are suitable mechanisms to allow attention shift.

Itti *el al.* have proposed a complete implementation of the saliency-based model in [4] and this model has been widely used by many researches related to visual attention. The model is again shown in Figure 2-11 for reference. In the following subsections, we will briefly explain the detail of this model.

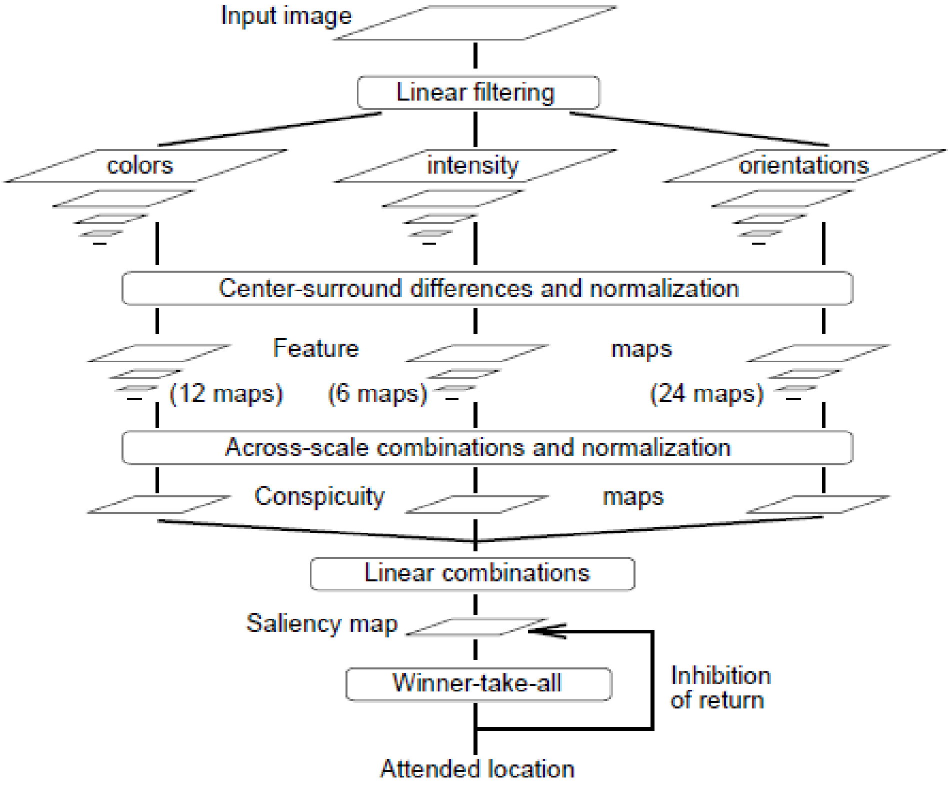


Figure 2-11 *General architecture of the model proposed by [4].*

2.3.1. Extraction of Early Visual Features

Given an input image, the first processing step consists of decomposing this input into a set of distinct “channels,” by using linear filters tuned to specific stimulus dimensions, such as luminance, red, green, blue and yellow hues, or various local orientations.

- Intensity: With r, g and b being the red, green and blue channels of the input image, an intensity image *I* is obtained as $I = (r + g + b)/3$.

- Color: The r, g and b channels are normalized by I in order to decouple hue from intensity. However, because hue variations are not perceivable at very low luminance (and hence are not salient), normalization is only applied at the locations where I is larger than 1/10 of its maximum over the entire image (other locations yield zero r, g and b). Four broadly-tuned color channels are created: R = r – (g + b)/2 for red, G = g – (r + b)/2 for green, B = b – (r + g)/2 for blue, and Y = r + g – 2(|r - g| + b) for yellow (negative values are set to zero).

- Orientation: Four local orientation features according to the angles $\theta \in \left\{0^\circ, 45^\circ, 90^\circ, 135^\circ\right\}$ are used. Gabor filters, which represent a suitable mathematical model of the receptive field impulse response of orientation-selective neurons in primary visual cortex [9], are used to compute the orientation features.

After each channel feature maps have been obtained, their different spatial scales are created using Gaussian pyramids, which progressively low-pass filter and sub-sample the input image. In the implementation, the pyramids have a depth of 9 scales, like in Figure 2-12. The original image is shown on the top of Figure 2-12, which is treated as Level 0 in the pyramid. Each subsequent level is obtained by low-pass filtering and down-scaling its previous level by a factor 2 in the horizontal and vertical directions, respectively.

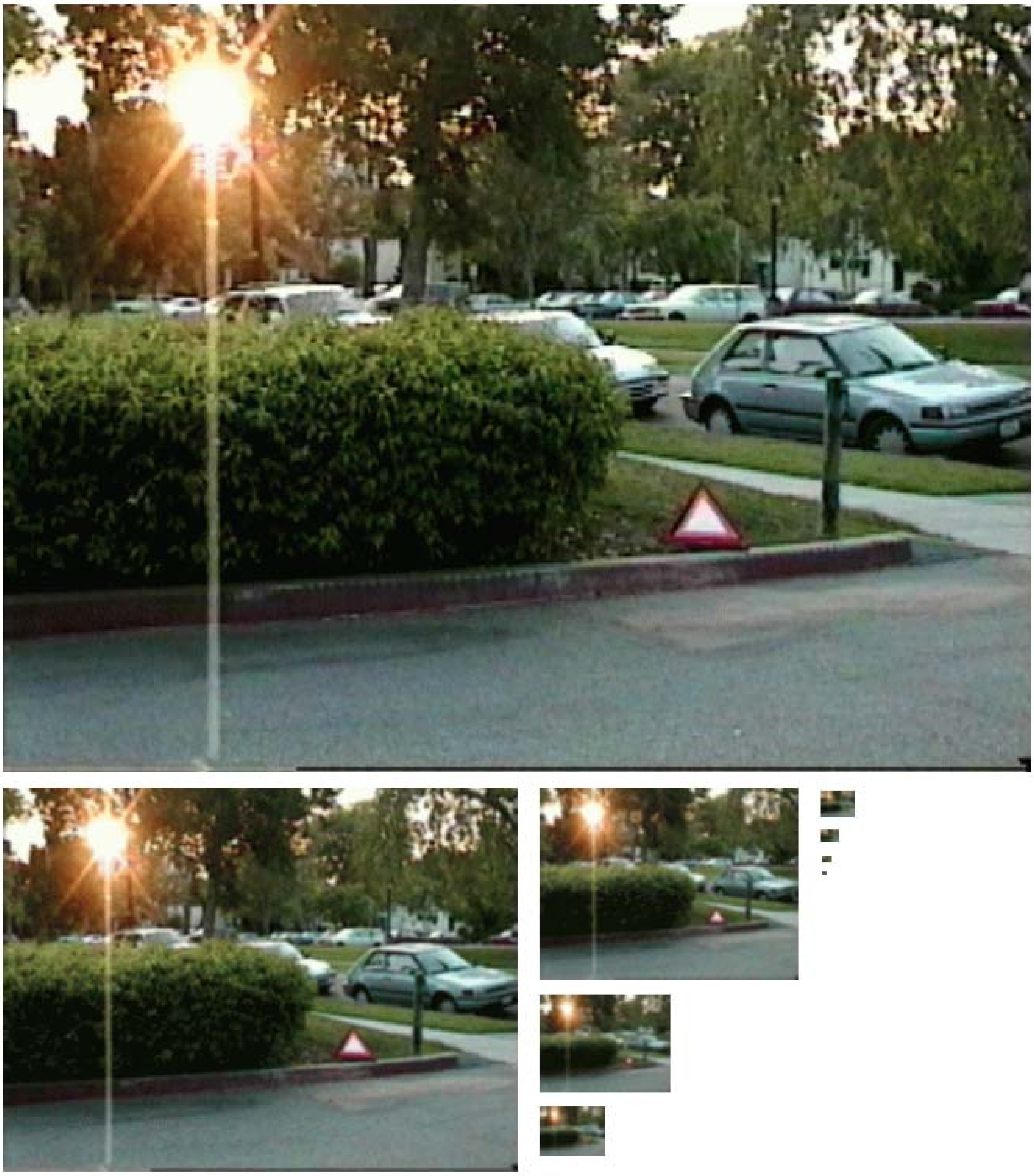


Figure 2-12 *Pyramidal representation of the input image [10].*

Next, each feature is computed in a center-surround structure similar to visual receptive fields. Using this biological paradigm makes the perception system sensitive to local spatial contrast, rather than amplitude, in that channel. The center-surround operation is implemented in the model as the difference between a fine scale and a coarse scale for a given feature.

2.3.2. Combining Information across Multiple Maps

For each feature map created from center-surround operations, the multi-scale maps are combined in a competitive way to form a unique feature-related conspicuity map. A combination of the feature maps provides the bottom-up input to the saliency map. At each spatial location, activity from these feature maps consequently needs to be combined into a unique scalar measure of salience.

However, because of the large number of maps being combined, the system faces a severe signal-to-noise ratio problem. A salient object may only elicit a strong peak of activity in one or a few feature maps, tuned to the features of that object. However, the combination of a larger number of feature maps may cause strong peaks at numerous locations. In order to solve such problems, [11] have proposed two feature combination strategies: contents-based global non-linear amplification, and iterative localized interactions.

2.3.2.1. Contents-based Global Amplification (N_1(.))

Given conspicuity maps, which should be integrated into a unique map, the normalization strategy $N_1(:)$ consists of the following steps:

1. Scale all maps to the same dynamic range in order to eliminate across-modality amplitude difference due to dissimilar extraction mechanisms.

2. For each map, compute the global maximum *M* and the average *m* of all the other local maxima. A local maximum of a map is defined as a location whose value is larger than those of its adjacent neighbors.

3. Globally multiply the map by a weight $\omega_M = (M - \overline{m})^2$. $N_1(.)$ normalizes a conspicuity map M in accordance with $N_1(M) = \omega_M \cdot M$.

In fact, ω measures how the most active locations differ from the average of local maxima of a conspicuity map. Hence, this normalization operator promotes the conspicuity maps in which a small number of strong peaks of activity are present. Maps that contain numerous comparable peak responses are demoted. This effect is clearly illustrated in Figure 2-13. The intensity map contains comparable responses which lead to a small ω. For this reason the intensity map is strongly suppressed. Due to the presence of a distinctive location in the orientation map, the corresponding ω is large and this explains the global amplification of that map. It is obvious that this competitive mechanism is purely data-driven and does not require any a priori knowledge about the analyzed scene.

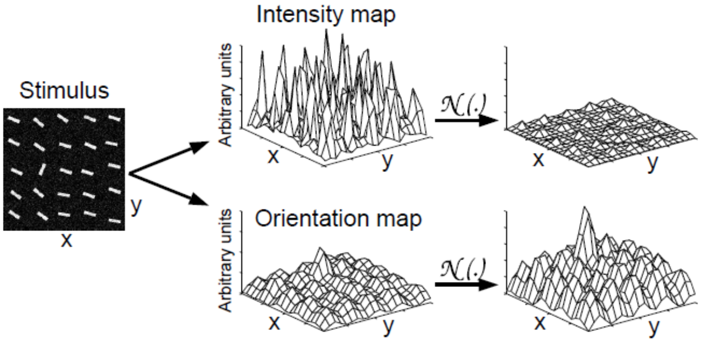


Figure 2-13 *Contents-based global amplification normalization [4].*

2.3.2.2. Iterative Non-Linear Normalization (N_2(.))

The non-linear normalization strategy N_2(.) is composed of the following steps. First, all maps are normalized to the same dynamic range in order to remove modality-dependent amplitude differences. Second, each map is iteratively convolved by a large 2D *DoG* (Difference of Gaussian) filter. The negative results are clamped to zero after each iteration. At each iteration of the normalization process, a given map *M* is transformed in accordance with Eq. 2-1.

$$M \leftarrow \left|M * Dog\right|_{\geq 0}, \qquad \textit{Eq. 2-1}$$

where $(*)$ is the convolution operator and $\left|.\right|_{\geq 0}$ discards negative values.

The normalization strategy N_2(:) relies on simulating local competition between neighboring conspicuous locations. Spatially grouped locations, which have similar conspicuities, are suppressed; whereas spatially isolated conspicuous locations are promoted. The behavior of the iterative non-linear normalization method is illustrated in Figure 2-14. The upper example of Figure 2-14 illustrates how the non-linear normalization progressively promotes the major peak while suppressing less conspicuous locations. On the contrary, the bottom example of Figure 2-14 shows the suppressing of the entire map in the absence of prominent peaks.

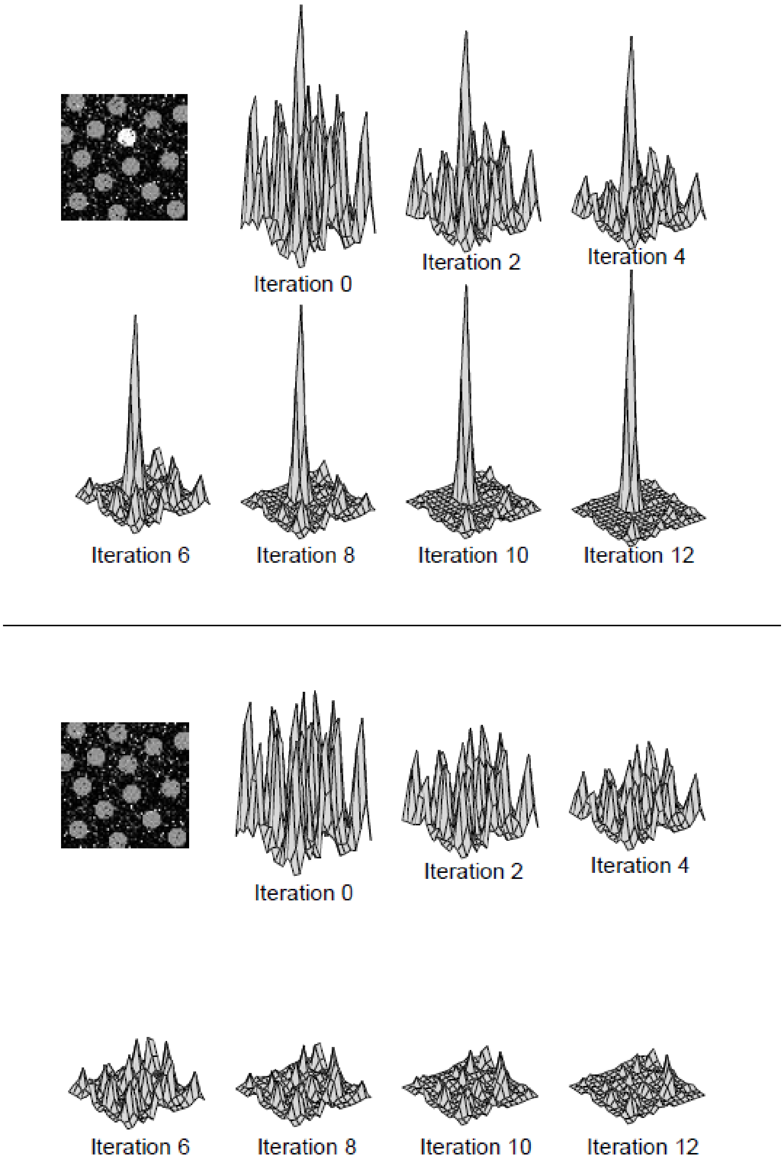


Figure 2-14 *Iterative non-linear normalization [11].*

Besides being inspired from the human vision [11], this normalization strategy, thanks to its non-linearity, has the advantage of noise suppression while promoting the major peaks in the conspicuity map.

After the conspicuity maps of each channel have been obtained, the saliency map is formed by averaging these three conspicuity maps. Ffinally, the winner-take-all and inhibition of return processes are applied on the saliency map to achieve selective attention.

2.4. INTENSITY-PAIR DISTRIBUTION

In [12], Jen *et al*. proposed an intensity-pair distribution technique, which was used to enhance image contrast. This distribution possesses both local information and global information of the image content. For a given image, this method tests at each pixel the intensity difference between that pixel and each of its 8-connection neighbors. Figure 2-15 shows an illustration of a pixel and its 8-connection neighbors. Due to the commutative property of intensity pair, we only check 4 neighboring pixels, instead of 8, as we scan the image in the raster order. That is, for the pixel at E in Figure 2-15, we only check the intensity difference between that pixel and its upper-left pixel (A), upper pixel (B), upper-right pixel (C), and left pixel (D) [12].

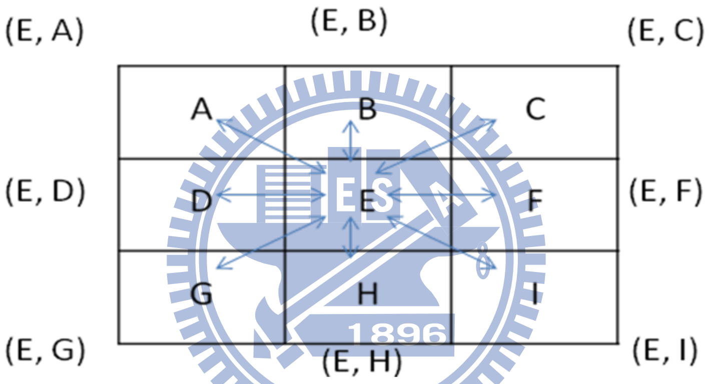
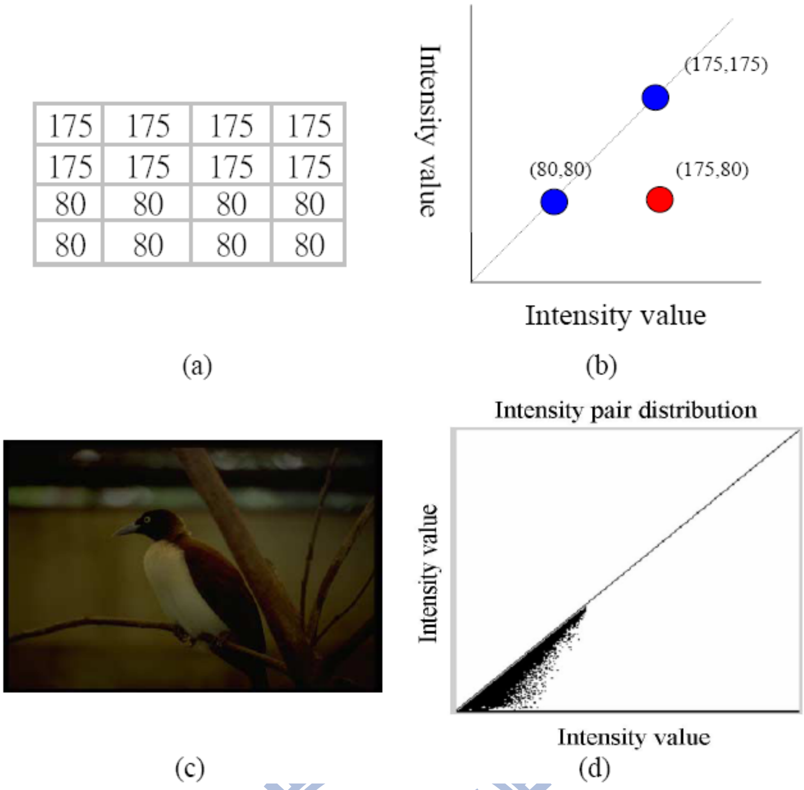


Figure 2-15 *An illustration of a pixel and its 8-connection neighbors.*

After the computation of intensity differences, we may imagine that we have formed an intensity-pair distribution as shown in Figure 2-16. Figure 2-16(a) shows an example of a 2-D image. As we calculate the intensity difference between adjacent pixels, we form four different intensity pairs, {(80, 80), (175, 80), (80, 175), (175, 175)}. If we ignore the pair order and treat (175, 80) and (80, 175) as the same type of pair, these four types of pairs are further merged into three types of pairs, {(80, 80), (175, 80), (175, 175)}. As we count the total pixel number for each type of intensity pair, we may generate the intensity-pair distribution as shown in Figure 2-15(b). Here, the values at (80, 80), (175, 80), and (175, 175) are 21, 13, and 21, respectively. Similarly, for a real image shown in Figure 2-16(c), its intensity-pair distribution can be easily calculated as shown in Figure 2-16(d). Especially, if the intensity difference of an intensity-pair is large than a pre-selected threshold, that intensity-pair is treated as an edge pair [12].



175	175	175	175
175	175	175	175
80	80	80	80
80	80	80	80

(a) (b)

(c) (d)

Figure 2-16 *(a) A synthesized image (b) Intensity-pair distribution of (a) (c) A real image (d) Intensity-pair distribution of (c) [12]*

By analyzing the content of intensity-pair distribution, we can get useful information for the detection of visual salient region. This will be discussed in later chapter.

Chapter 3.

PROPOSED METHOD

The goal of saliency map is to capture the regions where a person may pay more attention to. As mentioned earlier, bottom-up methods are more flexible and are applicable to different scenarios. However, the major problems of bottom-up visual saliency models are their complicated models and the difficulties in detecting and labeling regions in complex natural scenes. In a bottom-up approach, we aim to detect those regions which are “special” or “abnormal”. In this thesis, we develop our system based on the following two intuitive assumptions:

(1) A region with a strong contrast with respect to its surrounding regions is more likely to be paid attention to.

(2) A region is less attractive to the observer if its property is common in the scene.

With these two assumptions, we develop our saliency region detector based on the infrastructure proposed by Itti [4]. The flow chart of the proposed saliency region detector is illustrated in Figure 3-1. In the following sub-sections, we will explain in detail the sub-modules of this system.

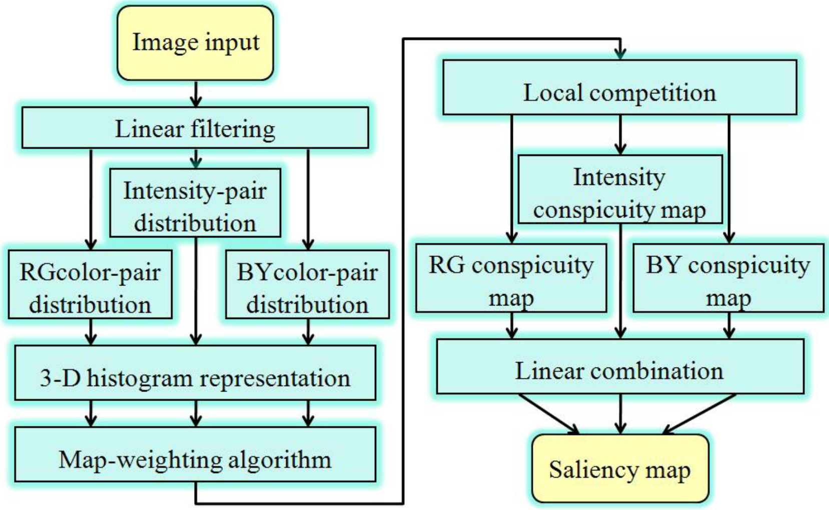


Figure 3-1 *Block diagram of the proposed system*

3.1. LINEAR FILTERING OF IMAGE DATA

Similar to Itti's approach, we decompose an input image into a few feature vectors, including intensity, RG color, and BY color. Here, we ignore the orientation feature since the orientation feature is usually not a dominating factor in natural scenes. In our system, the intensity channel is defined as:

$$I = \frac{(r+g+b)}{3} \tag{Eq. 3-1}$$

where r, g, and b denote the red, green, and blue components of the input image.

On the other hand, we define the red, green, blue, and yellow hues of the image pixel as:

$$R = r - \frac{(g+b)}{2} \tag{Eq. 3-2}$$

$$G = g - \frac{(r+b)}{2} \tag{Eq. 3-3}$$

$$B = b - \frac{(r+g)}{2} \tag{Eq. 3-4}$$

$$Y = r + g - 2(|r-g| + b) \tag{Eq. 3-5}$$

For each color hue, negative values are set to zero. Each color hue yields the maximal response for the pure, fully-saturated hue and yields zero response for gray colors. These four color hues are then merged together to form two opponent-color channels that mimic the color opponent process in human's visual system [8].

Since the separated color feature maps have obtained, we are going to introduce why and how to combine them. It must refer to the biological functionality of human brain. In human brain, there exists a 'color opponent-component' system. In the center of receptive fields, neurons which are excited by one color (eg. Red) while inhibited by another color (eg. Green). Red/green, green/red, blue/yellow and yellow/blue are color opponent pairs which exists in human visual cortex [13]. Thus, in our approach, we define the RG color channel to be

$$RG = |R - G| \qquad \textit{Eq. 3-6}$$

and the BY channel to be

$$BY = |B - Y| \qquad \textit{Eq. 3-7}$$

Currently, we have three feature maps extracted from input image, as shown below:

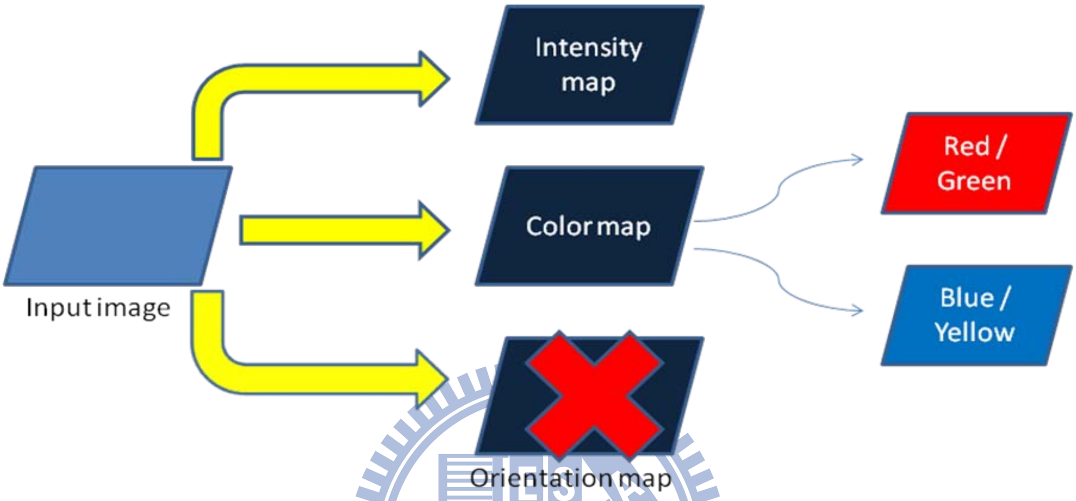


Figure 3-2 *A sketch diagram of low-level image feature extraction*

These two opponent-color channels, together with the I (Intensity) channel, are fed into the following modules to form feature-pair distributions.

3.2. FEATURE-PAIR DISTRIBUTIONS

For each of the I, RG, and BY channels, we compute the feature-pair distribution as proposed in [12]. As mention in Section 2.4, Jen *et al*. proposed the concept of intensity-pair distribution for the enhancement of image contrast. Since this distribution possesses both local information and global information, it may offer useful information for us to detect visual saliency regions in the image. In a feature-pair distribution, the global information tells us what kinds of features are common in the image, while the local information tells us which portions of the image may exhibit large contrast. Hence, by properly using the pair-distributions of the I, RG, and BY channels, we can efficiently detect those image portions with unusual appearance or with stronger contrast.

To establish the feature-pair distribution for the I channel, we check at each pixel the intensity pairs between that pixel and its 8-connection neighbors. Figure 2-15 shows an illustration of a pixel and its 8-connection neighbors. If we denote the I values of these nine pixels as A to I, respectively, then the eight intensity pairs {(E, A), (E, B), (E, C), (E, D), (E, F), (E, G), (E, H), and (E, I)} are formed and accumulated in the feature-pair distribution. Clearly, we can expect that the intensity pairs over smooth regions will lie around the 45-degree line; whereas these intensity pairs across edges will lie somewhere away from the 45-degree line.

Figure 3-3 shows an example of the intensity-pair distribution. For the airplane image shown in Figure 3-3, since the sky and grass are the major backgrounds of the image, the intensity pairs over these two regions form two major clusters in the intensity-pair distribution. Here, we intentionally colorize these two clusters to indicate their correspondence. On the other hand, the aircrafts map to a smaller cluster in the lower-left corner of the distribution. Moreover, the intensity pairs over the sky-grass boundary and the aircraft-sky boundary form four clusters (represented in red color) far away from the 45-degree line. Based on this intensity-pair distribution, we can easily deduce that the boundary between the aircraft and the sky exhibits a stronger contrast than the sky-grass boundary. With the facts that (1) the aircraft is “less common” than the sky and the grass; and (2) the aircraft has a stronger contrast with respect to its background, we may deduce that these two aircrafts may catch the attention of most observers.

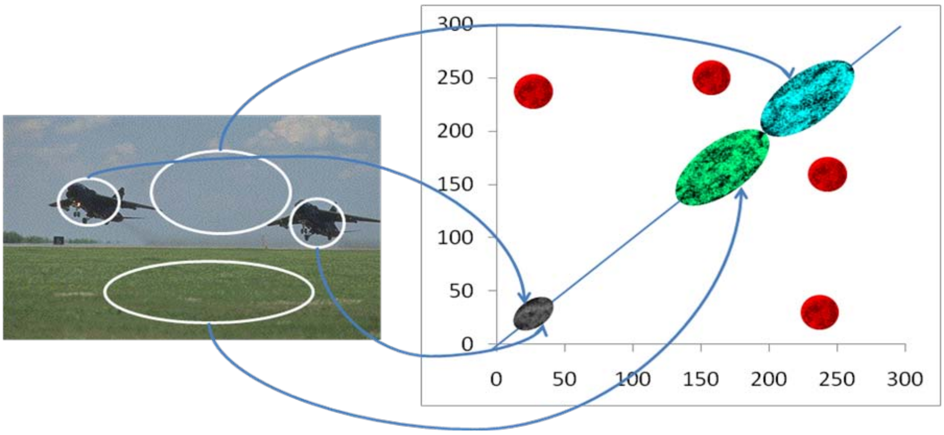
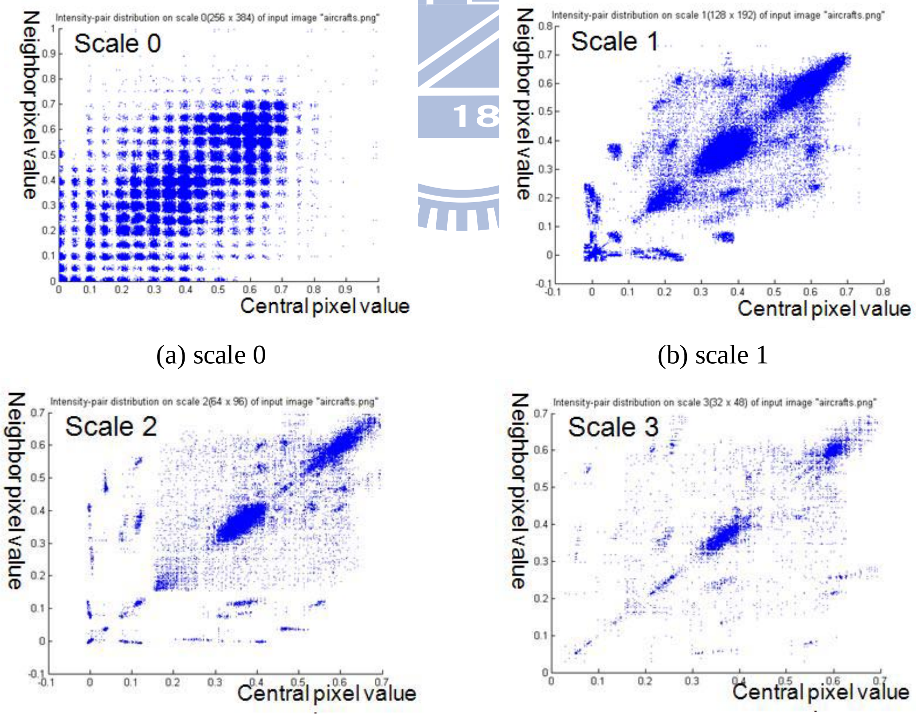


Figure 3-3 *A matching example of modified intensity-pair distribution*

Here rises a question: how large should the input image be? In Figure 3-4, we show four intensity-pair distributions with their input image being scale 0 to scale 3. When the scale is increased by 1, the image’s height and width are reduced by 2, respectively. The choice of scale is image dependent. However, in Scale 0 or Scale 1, the image usually contains quite a large number of scattered data and requires longer processing time. Hence, in our approach, we typically work on Scale 2 and Scale 3, as shown in Figure 3-4(c) and (d).



(a) scale 0

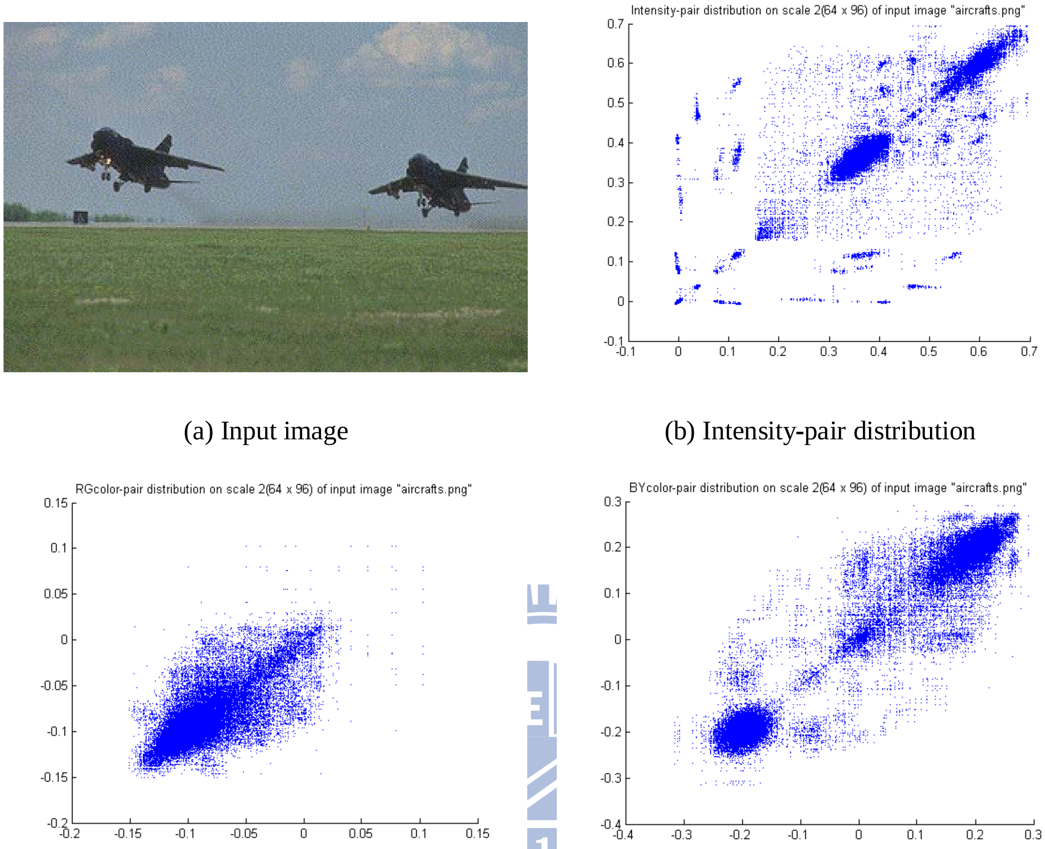
(b) scale 1

(c) scale 2

(d) scale 3

Figure 3-4 *An example of intensity-pair distribution with different scale input*

Based on the same concept, we can form the RG-pair distribution for the RG channel, and the BY-pair distribution for the BY channel. These three feature-pair distributions may offer us plentiful clues about the global statistics and the local variations of the image contents.



(a) Input image

(b) Intensity-pair distribution

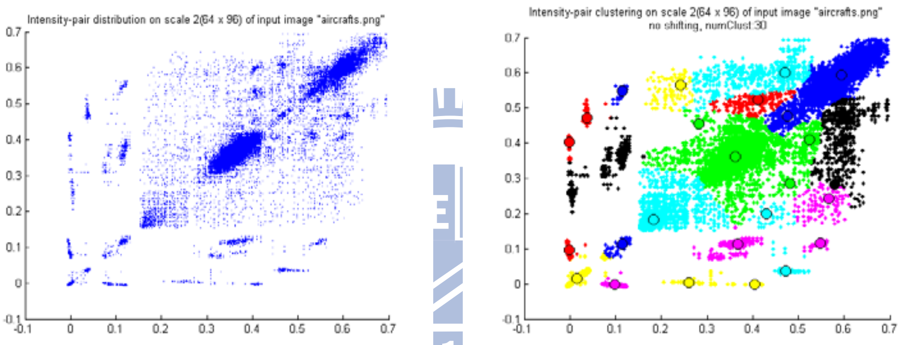
(c) RG color-pair distribution

(d) BY color-pair distribution

Figure 3-5 *An example of the feature-pair distributions*

3.2.1. CLUSTERING

To identify the most common properties in the image, we need to identify the major clusters in the feature-pair distributions. From the feature-pair distributions obtained at the previous section, there are apparent clusters which we can tell easily. The existing clustering algorithms seem to be a good tool for us to segment each cluster out. Figure 3-6 is an example of the intensity-pair distribution processed by the mean-shift clustering algorithm. The resulting clusters are reasonably good. Unfortunately, these existing clustering algorithms are usually computationally expensive and time-consuming. These disadvantages disobey our major requirement that the system should not possess complicated computations and should be fast enough for real-time processing and analysis.



(a) (b)

Figure 3-6 *An example of intensity-pair distribution after mean-shift clustering (a)intensity-pair distribution (b) mean-shift clustering algorithm passing through (a)*

3.2.2. 3-D Histogram Representation

To simplify the computations, we choose another approach that operates over the feature-pair distributions directly. In Figure 3-7, we show the 3-D histogram representation of the feature-pair distribution. This 3-D histogram is formed by dividing the x-y plane into a few uniform cells and count for each cell the total number of feature pairs within that cell. Clearly, we can expect that, in general, most clusters occur around the diagonal line in the 3-D histogram since most regions in a natural image are smooth. Moreover, the background elements would yield the largest cluster since the background usually occupies the largest area in the image. On the contrary, foreground objects usually correspond to smaller clusters. Besides, those clusters away from the diagonal line correspond to the boundary regions or the texture regions in the image.

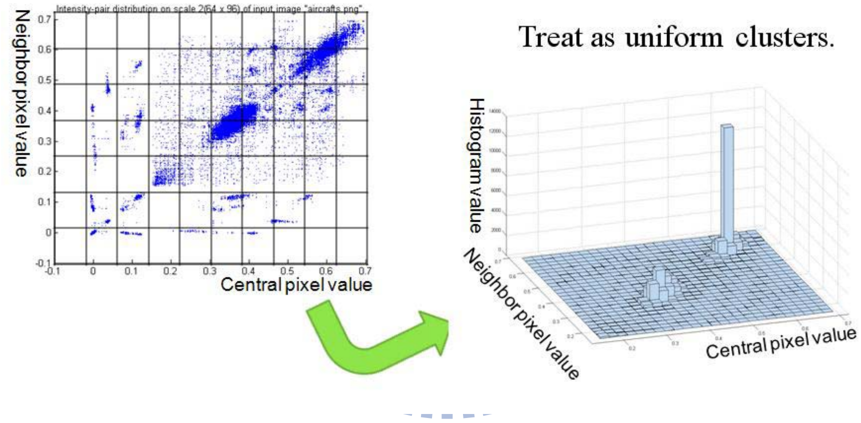


Figure 3-7 *The 3-D histogram representation of feature-pair distribution.*

In this 3-D histogram, we denote the cell at the intersection of the ith column and jth row as C(i,j). We further define a cell to be a “diagonal” cell when $|i-j| \leq D_{th}$, where D_{th} is a pre-selected threshold. On the contrary, a cell is defined as “off-diagonal” if $|i-j| > D_{th}$.

3.3. Map-Weighting Algorithm

One of the reasons why we use image feature-pair distributions is that we may perform straightforward weighting strategy on them to form saliency map. Clearly, we can expect that, in general, the background elements would yield the largest clusters since backgrounds always occupy the most part of the scene in the spatial domain. On the contrary, foreground objects usually occupy a smaller space in an image. Hence, foreground objects, or salient regions, will form smaller clusters in opposition to the background. As for the clusters away from 45-degree line, it is apparently that they represent edge clusters, since they have a strong difference between the central pixel and the neighboring pixels.

In our approach, without using clustering algorithm, we build a map-weighting algorithm to directly weigh the saliency degree of the cells in the 3-D histogram. This weighting algorithm contains two main parts: the "contrast weight" to gather the information concealed in off-diagonal cells; and the "distinction weight" that determines how likely a diagonal cell may contain the intensity pairs from a visually salient region. A simple structure of the algorithm is shown below:

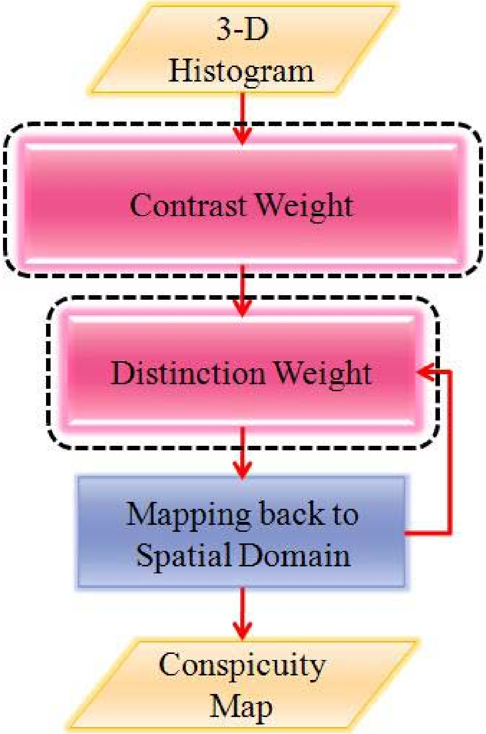


Figure 3-8 *A simple structure of map-weighting algorithm*

3.3.1. CONTRAST WEIGHT

As mentioned above, these clusters over the off-diagonal cells correspond to the boundary portions or texture portions in the image. As an off-diagonal cell is far away from the diagonal line, it indicates that any intensity pair within this cell will exhibit stronger contrast. Intuitively, we may use this kind of information to estimate how likely a region may attract observers’ attention.

Here, we give an example to explain the calculation of this contrast weight. Given a smooth region R_0 with the feature value f_0, this region would correspond to a cluster in the cell that contains the feature pair (f_0, f_0). If this smooth region has a surrounding region R_1 with the feature value f_1, we expect there is a cluster at the cell containing the pair (f_0, f_1) and a cluster at the cell containing the pair (f_1, f_0). If these two cells are far away from the diagonal line, then there should be a strong contrast between R_0 and R_1. Moreover, if these two off-diagonal cells contain a large number of feature pairs, it means R_0 may share a long boundary with R_1.

Figure 3-9 illustrate the concept of contrast weight. In Figure 3-9, lines a to d represent the four different profiles of the pair-distribution map in the left of Fig 3-9. In each profile, there is a diagonal cluster, as represented by the yellow-green block, together with several off-diagonal clusters. As an off-diagonal cluster is far away from the diagonal line, we assign a larger weight for it, as represented by the light blue curves in Figure 3-9.

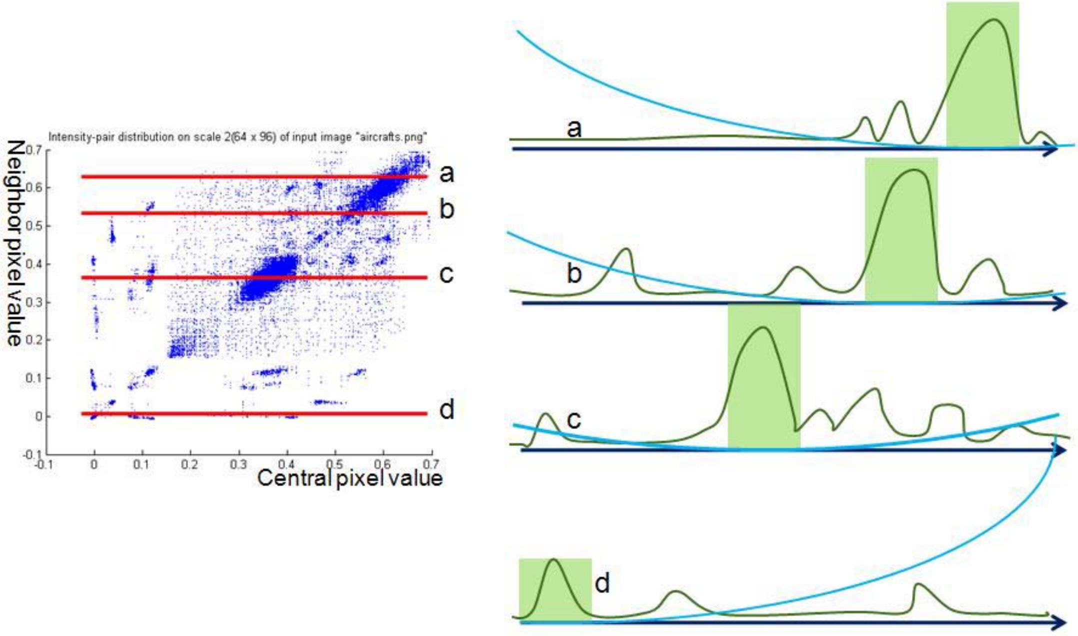


Figure 3-9 *Illustration of concept of contrast weight*

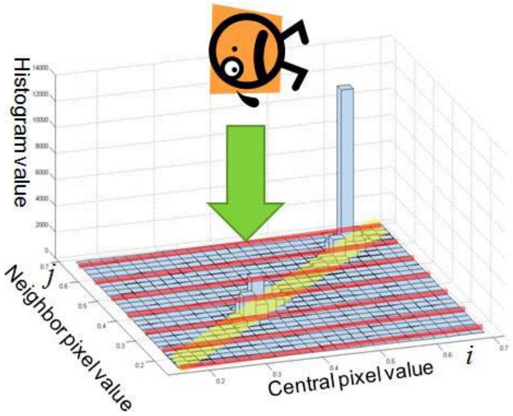


Figure 3-10 *Viewpoint of contrast weight*

Hence, given an off-diagonal cell C(i,j), we define its self-contrast-weight as

$$self\_contrast\_weight(i,j) = hist(i,j) \times |i-j|^2 \qquad \text{Eq. 3-8}$$

where hist(i,j) denotes the total number of intensity pairs in the cell C(i,j). Here, we take the square of |i-j| to emphasize those cells far away from the diagonal line.

With the definition of self-contrast-weight for off-diagonal cells, we further define the contrast-weight for diagonal cells, which are defined as C(i,j) with |i-j| ≤ D_{th}. In our algorithm, D_{th} is chosen to be a small-value constant. Here, for a diagonal cell at C(i,j), we define its contrast weight as

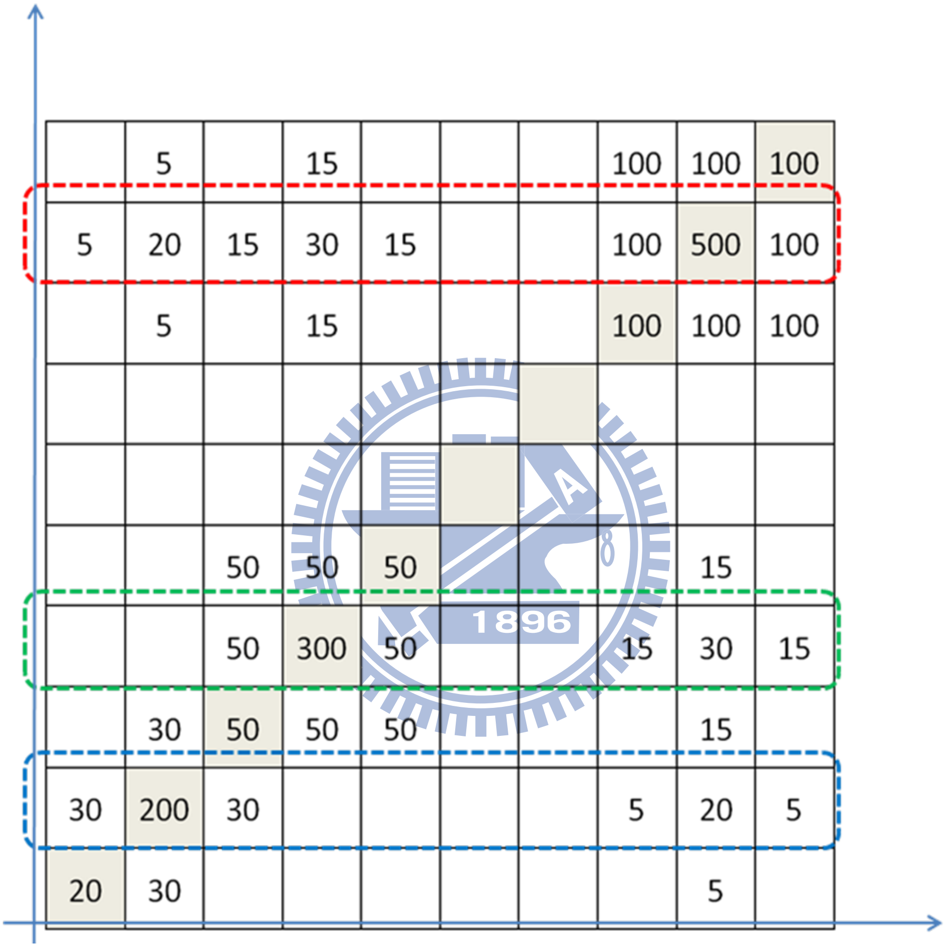
$$\begin{aligned} &contrast\_weight(i,j) \\ &= \sum_{\forall k} self\_contrast\_weight(i,k) \end{aligned} \qquad \text{Eq. 3-9}$$

or

$$\begin{aligned} &contrast\_weight(i,j) \\ &= \sum_{\forall k} self\_contrast\_weight(k,j) \end{aligned} \qquad \text{Eq. 3-10}$$

That is, we sum up the self-contrast-weights for all the cells along the ith column or along the jth row.

From the 3-D histogram, we have the value of each square region, as shown in Figure 3-11. The numbers shown in each square represent the height, or the number of feature-pairs, in each histogram cell. Figure 3-11 illustrates an example to explain the calculation of contrast weight. In this case, we define D_{th} = 1. For each diagonal cell, we check its entire horizontal neighbors. For example, in Figure 3-11, the white cell in the red rectangular area has the contrast weight 2830, which is computed as $5\times8^2+20\times7^2+15\times6^2+30\times5^2+15\times4^2=2830$, and the white cell in the green area has the contrast weight $15\times6^2+30\times5^2+15\times4^2=1530$. Note that since D_{th} = 1 the cells next to the white cell are also considered as diagonal cells and are not included in the computation of contrast weight. Moreover, for these cells next to the white cells, their contrast weight are computed in the same way as that of the white cell. After every horizontal line is scanned, we get the contrast weight for every diagonal cell.



	5		15				100	100	100
5	20	15	30	15			100	500	100
	5		15				100	100	100
		50	50	50				15	
		50	300	50			15	30	15
	30	50	50	50				15	
30	200	30					5	20	5
20	30							5	

Figure 3-11 *Example of 2-D view of 3-D histogram*

3.3.2. Distinction Weight

After the estimation of contrast distribution, we further take into account the phenomenon that humans tend to pay less attention to the common regions in the image. Hence, beside the contrast weight, we add one more weight, named distinction weight, for these diagonal cells. This distinction weight is calculated in an iterative manner.

3.3.2.1. First Iteration

At the beginning, the diagonal cell with the largest value of *hist* is identified. Assume this cell is at C(i_1,j_1) and its *hist* value is denoted as hist(i_1,j_1). The distinction weight of this cell is defined as

□ *Eq. 3-10*

A sample 3-D histogram for the first iteration is shown in Figure 3-12, where max_hist = hist(i_1,j_1). This identified cell C(i_1,j_1) typically corresponds to the image background, the commonest portion of the image. Hence, by taking the reciprocal of hist(i_1,j_1), these background portions are assigned a lower value of distinction. That is, the commonest portion of the image is expected to be less visually salient to the observers.

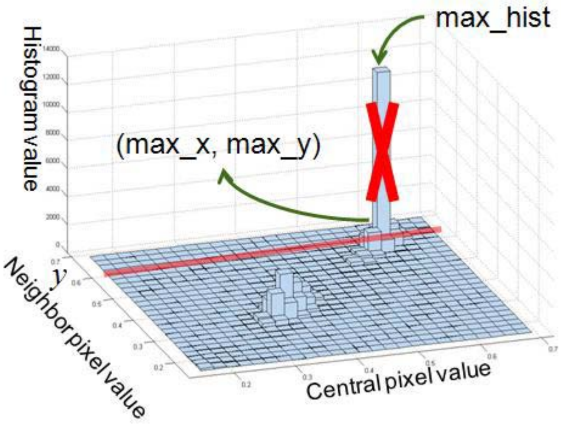


Figure 3-12 *An example of distinction weight for first iteration*

Since we have kept the coordinates of all the feature pairs within each cell, the weights of the diagonal cells can be mapped back to the spatial domain easily. At the same time, after the identification of the largest peak in the 3-D histogram, the value of the maximal peak is set to zero in order to run the next iteration.

3.3.2.2. THE SECOND ITERATION

After the largest cluster of the 3-D histogram being identified, we search for the second largest cluster. Assume the second largest cluster is identified at the cell C(i_2,j_2), its distinction weight is defined as

$$\begin{aligned} & distinction\_weight(i_2, j_2) \\ & = contrast\_weight(i_2, j_2) \times self\_distinction\_weight(i_2, j_2)’ \end{aligned} \qquad \text{Eq. 3-11}$$

where $$self\_distinction\_weight(i_2, j_2) = \frac{1}{hist(i_2, j_2)} \times d\,. \qquad \text{Eq. 3-12}$$

Here, d is the distance between (i_2,j_2) and (i_1,j_1). The inclusion of *d* in Eq. 3-12 is due to the reason that the regions corresponding to C(i_2,j_2) will not be visually salient to the observer if their feature values are too close to the feature value of the background. For example, in Figure 3-13, we get three candidate cells with the same number of feature-pairs. These three cells have three different distances, denoted as d1, d2, and d3, away from the largest peak in the histogram. Clearly, d3 is the farthest of the three. Hence, the order of the corresponding self_distinction_weights of these three candidate cells would be (3) > (2) > (1). In this case, we’ll pick the farthest cell as the second largest cell. In Figure 3-14, we illustrate the calculation of the distinction weight for the second largest cluster in the 3-D histogram.

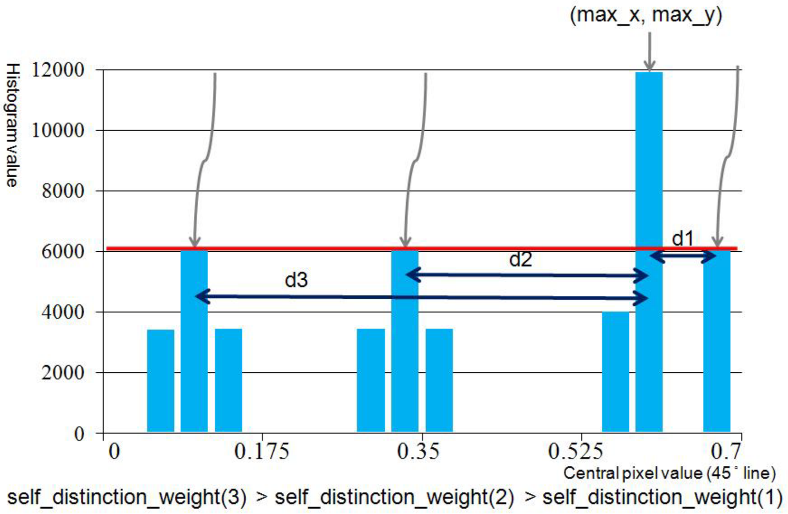


Figure 3-13 *Discussion of the influence of d*

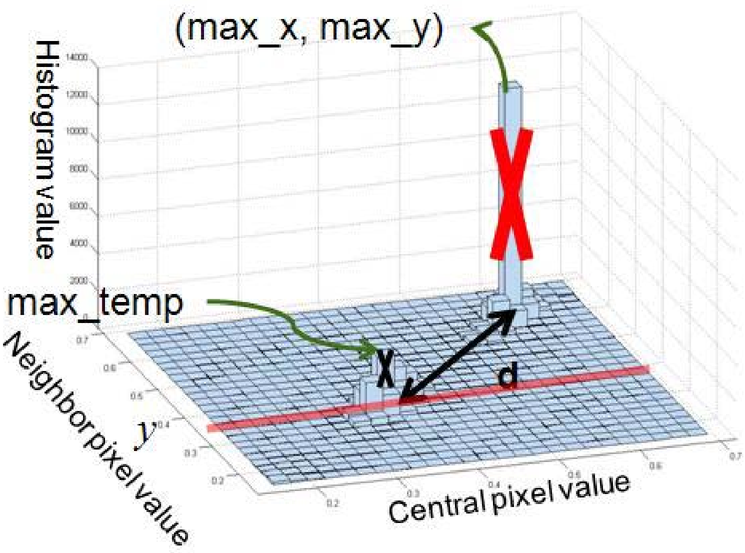


***Figure 3-14** An example of distinction weight for second iteration*

Similarly, since we have kept the coordinates of all the feature pairs within each cell, the weights of the diagonal cells can be mapped back to the spatial domain. After the identification of the second largest peak in the 3-D histogram, the value of that peak is set to zero in order to run the next iteration.

3.3.2.3. THE NEXT ITERATION

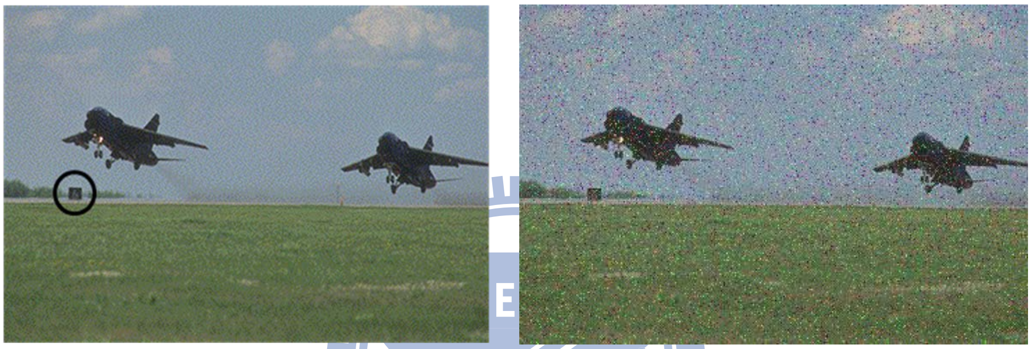
After the identification of C(i_2,j_2), we keep searching for the next largest cluster, C(i_3,j_3). For C(i_3,j_3), its distinction weight is defined as

$$\begin{aligned} &distinction\_weight(i_3, j_3) \\ &= contrast\_weight(i_3, j_3) \times \frac{1}{hist(i_3, j_3)} \times \sqrt{(i_3 - i_1)^2 + (j_3 - j_1)^2} \qquad \textit{Eq. 3-13} \end{aligned}$$

The same process is repeated until hist(i_k,j_k) is below a pre-defined threshold H_{th}, which is used to ignore small regions and suppress noise interference.

3.3.2.4. STOP CONDITION

Noise is always a key problem in image processing, so is in saliency region detection. If an input image is corrupted by noise, the noise might decrease the performance of salient detection result. As for trivial objects, such as the black-circle region in Figure 3-15(a), it is obvious that the trivial region will cause a quite small cluster in the feature-pair distribution. In order to suppress noise interference or avoid the detection of such trivial objects, a threshold should be chosen appropriately to stop the iterative search of histogram peaks. As an example shown in Figure 3-16, if we set the threshold H_{th} to 20, the histogram values below 20 will not go through the algorithm.



(a) Trivial region (b) With noise interference

Figure 3-15 *Sample highlight region of trivial region and noise interference*

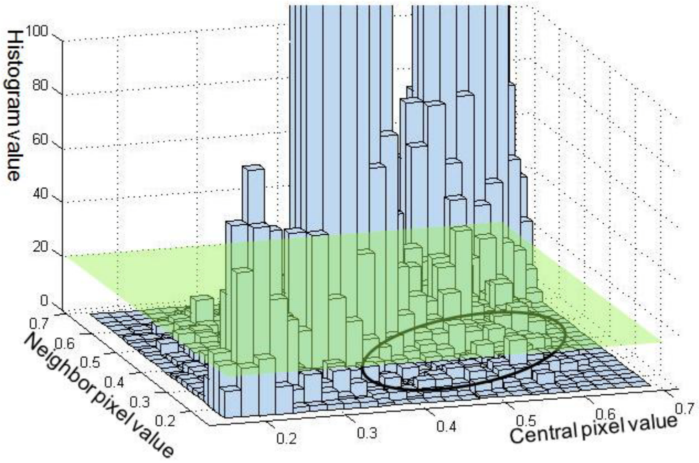
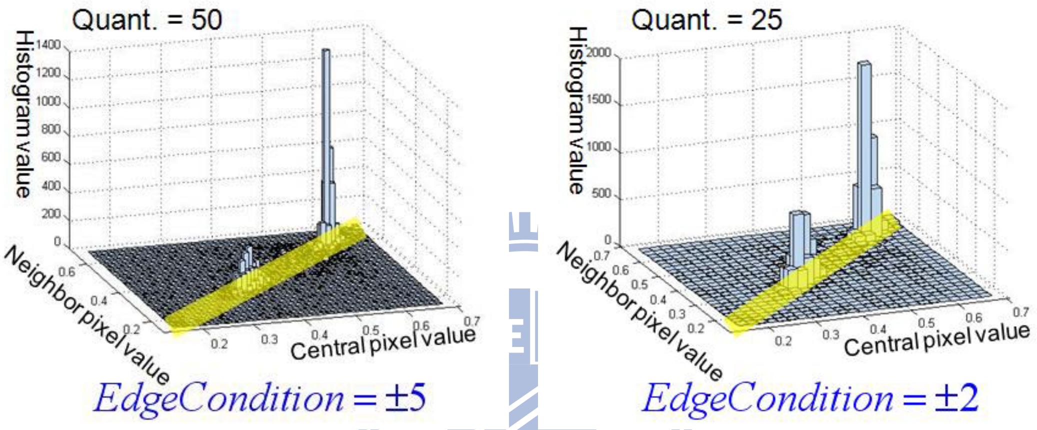


Figure 3-16 *An example of pre-define threshold H_{th} set to 20*

3.3.2.5. EDGE CONDITION

In the above procedure, we only search histogram peaks over diagonal cells and ignore all off-diagonal cells. Here, we define a cell C(i,j) to be a diagonal cell if $|i-j| \le D_{th}$. This is because an off-diagonal cluster is usually small and only corresponds to the boundary of some region in the image. Since we aim to detect visual saliency regions but not their boundaries, we only need to focus on diagonal cells but not off-diagonal cells. Here, we set a threshold D_{th} to determine how far away a cell may depart from the 45-degree line if that cell is to be treated as a diagonal cell. In Figure 3-17, we illustrate the range of diagonal cells in yellow for two different cases.



(a) Histogram quantization with 50 (b) Histogram quantization with 25

***Figure 3-17** Examples of two histograms with different quantization levels*

Actually, the threshold D_{th} depends on how we divide the domain of the 3-D histogram. If we divide the domain into 50 by 50 cells, as shown in Figure 3-17(a), we may choose a wider range for diagonal cells. For example, we choose $D_{th} = 5$ in Figure 3-17(a). On the other hand, if we only divide the domain of the 3-D histogram into 25 by 25 cells, we choose D_{th} to be a smaller value 2, as shown in Figure 3-17(b). In Figure 3-18, we show the detail flow chart of the whole map-weighting algorithm.

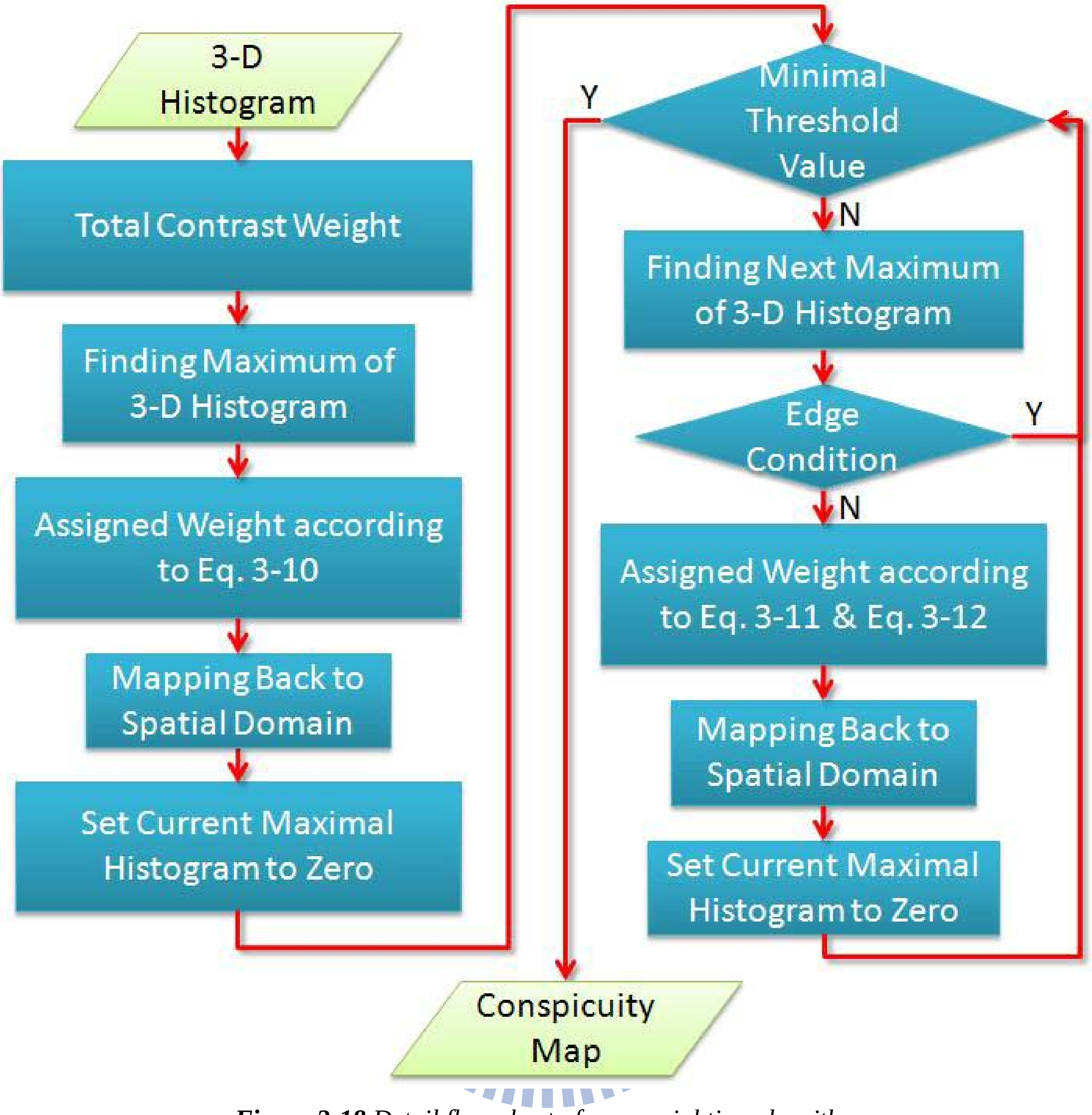

3-D Histogram
Total Contrast Weight
Finding Maximum of 3-D Histogram
Assigned Weight according to Eq. 3-10
Mapping Back to Spatial Domain
Set Current Maximal Histogram to Zero
Minimal Threshold Value
Y
N
Finding Next Maximum of 3-D Histogram
Edge Condition
Y
N
Assigned Weight according to Eq. 3-11 & Eq. 3-12
Mapping Back to Spatial Domain
Set Current Maximal Histogram to Zero
Conspicuity Map

Figure 3-18 *Detail flow chart of map-weighting algorithm*

3.4. Spatial Competition

For each feature-pair distribution, after the calculation of distinction weights for its diagonal cells, we can back project these distinction weights onto the spatial domain to get the corresponding conspicuity map. Since we only adopt simple rules in the calculation of distinction weights, the conspicuity maps usually suffer from poor signal-to-noise ratio. Hence, in the proposed approach, we adopt the local competition technique proposed in [14] to suppress unwanted regions and emphasize visually salient regions in the normalization stage.

As proposed by Itti *et al*. [14], they introduced three kinds of methods which were presented in Section 2.3.2. The first is the global maximum normalization which has been introduced in Section 2.3.2.1. This method normalized each map to a fixed dynamic range, and then summed all maps according to their global maximum and local maximums. The method was simple but has poor performance. The second suggestion was to learn linear map combination weights based on the expected targets provided by the system. Even though this method may improve the performance of detection greatly, it requires the target information in advance.

In this system, we use local competition as the normalization stage to inhibit unwanted regions and, at the same time, exhibit the true salient regions. The general principle of local competition is to apply self-excitation and neighbor-induced inhibition over each pixel in the conspicuity map. Here, a simple spatial competition scheme is used and the operation kernel is modeled as the DOG (Difference of Gaussians) pattern, which yields excitation around the center but induces inhibition from surrounding neighbors (see Figure 3-19(a)). This DOG kernel is expressed as

$$\begin{aligned} DOG(x,y) = {} & c_{ex}^2 \frac{1}{\sqrt{2\pi\sigma_{ex_x}^2}} e^{-\frac{x^2}{2\sigma_{ex_x}^2}} \frac{1}{\sqrt{2\pi\sigma_{ex_y}^2}} e^{-\frac{y^2}{2\sigma_{ex_y}^2}} \\ & - c_{inh}^2 \frac{1}{\sqrt{2\pi\sigma_{inh_x}^2}} e^{-\frac{x^2}{2\sigma_{inh_x}^2}} \frac{1}{\sqrt{2\pi\sigma_{inh_y}^2}} e^{-\frac{y^2}{2\sigma_{inh_y}^2}} \end{aligned} \qquad \textit{Eq. 3-14}$$

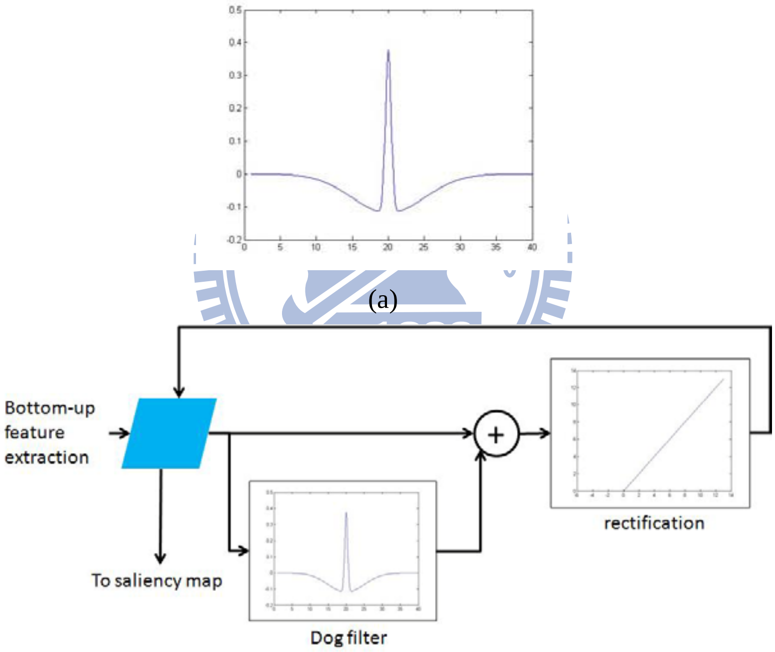
In our implementation, σ_{ex_i} and σ_{inh_i} are chosen to be the 2% and 25% of the image dimensions. For example, for a 100 by 80 image, we choose $\sigma_{ex_x} = 2$, $\sigma_{ex_y} = 1.6$, $\sigma_{inh_x} = 25$, and $\sigma_{inh_y} = 20$. On the other hand, we choose

$c_{ex} = 0.5$ and $c_{inh} = 1.5$.

In the local competition process, the three conspicuity maps are first normalized to the fixed dynamic range [0,1] to eliminate the factor of unequal dynamic ranges in different conspicuity maps. Each normalized conspicuity map is convolved with the 2-D DOG filter. The filtered result, together with a small negative constant, is added to the original map to form a new map. In the new map, all negative values are set to zero while the non-negative values are kept the same, as illustrated in Figure 3-19(b). This operation can be expressed as

$$M \leftarrow \left| M + M * DOG - K_{inh} \right|_{\geq 0} \qquad \textit{Eq. 3-15}$$

In our implementation, $\mathrm{K}_{\mathrm{inh}}$ is chosen to be 0.02. The introduction of this constant is to increase the speed of convergence over uniform texture regions.



(a)

(b)

Figure 3-19 *(a) A cross-section of DOG kernel. (b) System flow of spatial competition.*

For each conspicuity map, the above spatial competition process is repeated several times. If we choose a large number of iterations, the conspicuity map may converge towards the map of a single peak. On the contrary, if we choose a small number of iterations, the conspicuity map may still suffer from poor signal-to-noise ratio. In our implementation, we repeat 10 times the spatial inhibition process. Two examples of the time evolution of this process are shown in Figure 2-14, which illustrates that 10 iterations may yield adequate distinction between the two examples.

3.5. Linear Combination

After local competition, these three conspicuity maps are linearly combined into a single saliency map. Here, we use two kinds of process to weight each feature.

3.5.1. Naïve Combination

The first is purely to average the three feature conspicuity maps to get the saliency map. That is,

$$Saliency\ \ Map = \frac{I.Map + RG.Map + BY.Map}{3}, \qquad \textit{Eq. 3-16}$$

where the numerator terms stand for the three features respectively. The step described above is quite simple. Moreover, since the three maps have already been clamped to the same criterion at the stage of competitive normalization, this combination process requires almost no computational effort.

3.5.2. Data Driven Combination

However, for an input image, what really determines the salient region might be only intensity, or colors. As for colors, some regions might be salient in the red/green channel, while others might be in the blue/yellow channel. Hence, we may choose an adaptive combination that changes the weight according to the image characteristics. Here, we perform a data-driven approach and the summation is based on the following formula:

$$\begin{aligned} & Saliency\ \ Map \\ & = \frac{I.Map \times \max\_Ihist + RG.Map \times \max\_RGhist + BY.Map \times \max\_RGhist}{\max\_Ihist + \max\_RGhist + \max\_RGhist} \end{aligned} \qquad \textit{Eq. 3-17}$$

In Eq. 3-17, *I.Map*, *RG.Map*, *BY.Map* denote the conspicuity maps of the I-channel, RG-channel, and BY-channel, respectively. Max_Ihist, max_RGhist, and max_BYhist denote the largest peaks in the corresponding 3-D histograms. Typically, if a channel possesses a large peak in its feature-pair distribution, that channel is dominated by a specific feature value and the “unusual” regions usually become more apparent in the conspicuity map. Hence, we assign a larger weight for this channel.

Chapter 4.

EXPERIMENTAL RESULTS

Both computer simulation and subjective experiment were performed to verify the performance of the proposed algorithm. In computer simulation, the proposed algorithm is coded in Matlab without code optimization, and is tested over a PC with Intel® Core™2 Duo CPU running at 3G Hz. In the subjective experiment, an eye tracker is used to record the eye fixation points of 20 subjects in viewing 30 sample images. Figure 4-1 shows the eye tracker which borrowed from Prof. Chen-Chao Tao of Department of Commutation and Technology, NCTU. As we can see from Figure 4-1, the eye tracker looks just like a normal LCD monitor. At the bottom of the monitor, there are infrared emitters and sensors. The eye tracker use infrared and near-infrared non-collimated light to create a corneal reflection (CR). By detecting the strong reflectance from the observer's pupils, the eye track may determine the observer's eyes and then deduce the gaze focus of the eyes. [15].

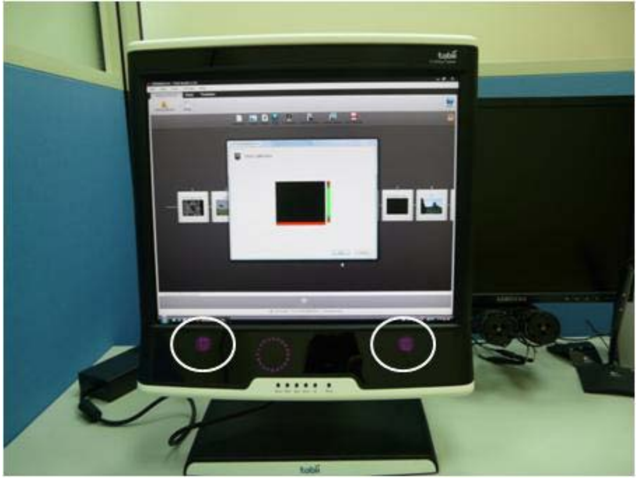


Figure 4-1 *The eye tracker we used for the experiment*

The subjects include both men and women. At the beginning of the experiment, all subjects were asked to sit comfortably on a chair and to glance freely at the popped out image. The distance between the subject and the screen is about 50 to 70 cm. Each image was shown only for 3 seconds to get the intuitive eye movement without concerning the internal state of each person. Between images, there was a 3-second short break. The eye fixation data of all these 20 subjects were averaged and compared with the results of computer simulation.

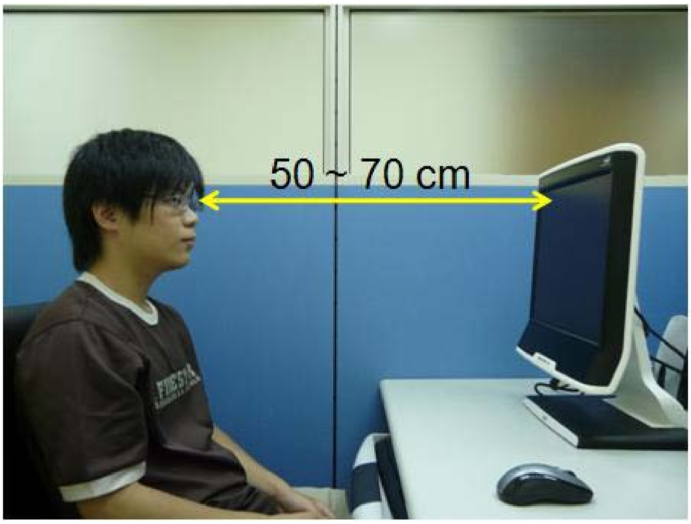


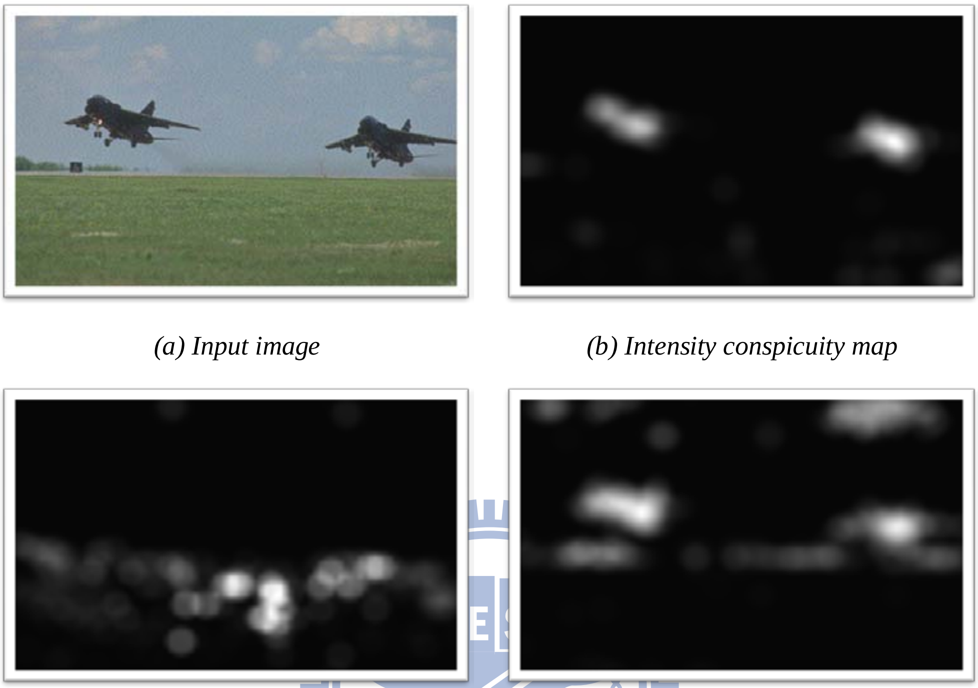
Figure 4-2 *Eye fixation experimental settings*

The computer simulation results of our technique are compared with human eye fixation data, which are extracted by averaging the eye fixation data of 20 subjects, together with the simulation results of two other algorithms. This comparison is to verify whether our method has the same, or even better performance if compared with other methods mentioned in Chapter 2. In the computer simulation, the parameter settings of our algorithm are listed in Table 4-1.

Table 4-1 *Test images and its parameter setting.*

	Scale	**Quantization**	**Stop condition**	**Edge condition**	**Execution time**
IMG – 1	2 (91 × 61)	25	30	±3	3.75 s
IMG – 2	2 (92 × 61)	25	30	±3	2.3 s
IMG – 3	2 (160 × 120)	25	30	±3	15.78 s
IMG – 4	2 (96 × 64)	25	30	±3	3.96 s
IMG – 5	3 (50 × 31)	25	50	±3	0.36 s
IMG – 6	2 (128 × 96)	25	30	±3	7.66 s
Comparison – 1	2 (100 × 62)	15	15	±5	2.66 s
Comparison – 2	1 (189 × 150)	15	5	±3	18.25 s
Comparison – 3	4 (80 × 77)	25	30	±5	4.52 s
Comparison – 4	2 (100 × 75)	25	30	±5	6.53 s

Figure 4-3 shows a sample input image and its three conspicuity maps, which are intensity in Figure 4-3(b), RG in Figure 4-3(c), and BY in Figure 4-3(d). From this three maps, we can see that the intuitive salient objects, the aircrafts, are popped out in Figure 4-3(b) and (c).

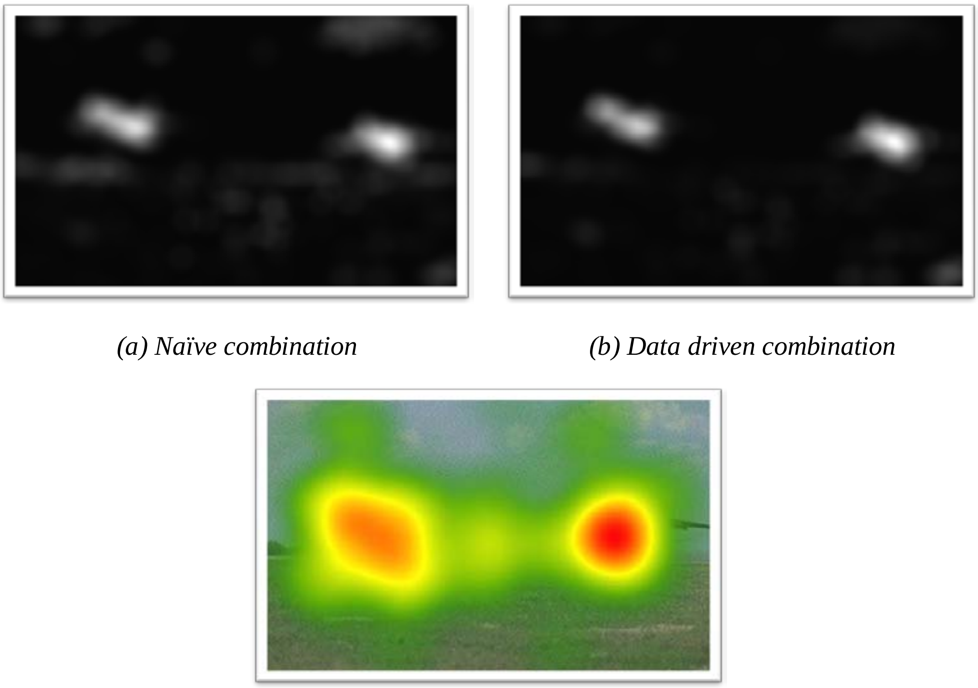


(a) Input image *(b) Intensity conspicuity map*

(c) RG color conspicuity map *(d) BY color conspicuity map*

***Figure 4-3** A sample input image and its three conspicuity maps*

After obtaining the three conspicuity maps as in Figure 4-3, the two combination strategies are used in order to see the difference between other. Figure 4-4(a) shows the resulting saliency map which is formed by the naïve combination; whereas Figure 4-4(b) is the result of the data-driven combination. From these saliency maps, the naïve combination yields more popped-out regions compared to the data-driven approach. In Figure 4-4(a), some unwanted regions appear which can be considered as noise interference. In Figure 4-4(b), the output map is more reliable and closer to the human eye fixation heat map.



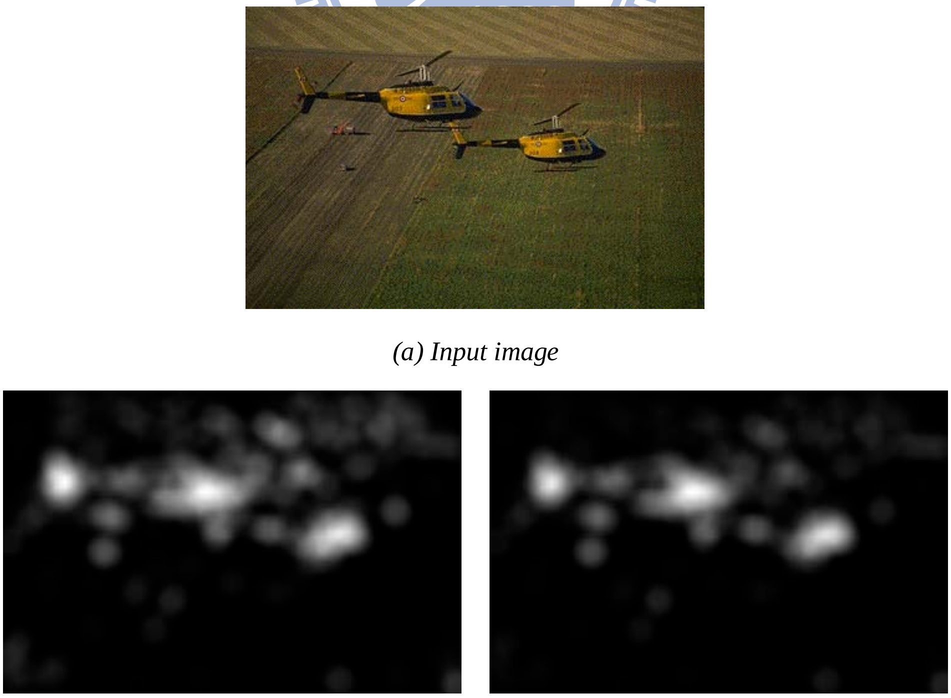
(a) Naïve combination

(b) Data driven combination

(c) Heat map from 20 subjects

Figure 4-4 *Resulting saliency maps of Figure 4-3 and heat map of human fixation (IMG - 1)*

Another clear result that shows the superiority of the data-driven combination is as follows.



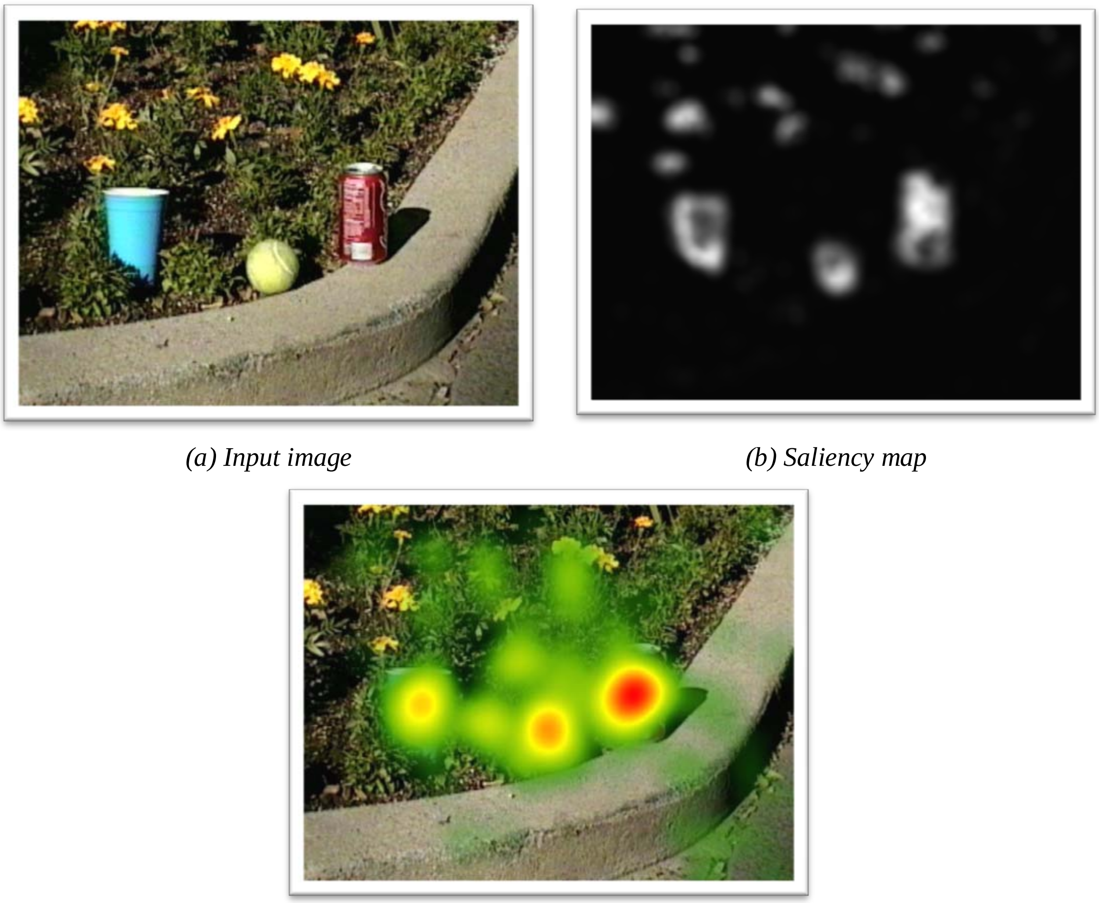
(a) Input image

(b) Naïve combination

(c) Data driven combination

Figure 4-5 *A more specific result explains the combination stage (IMG - 2).*

Based on the above discussion about the combination process, we thus use the data-driven method as the combination strategy in our saliency map detector. Figure 4-6 to Figure 4-9 show the experimental results for a few nature images. The human eye fixation heat map is presented for comparison.



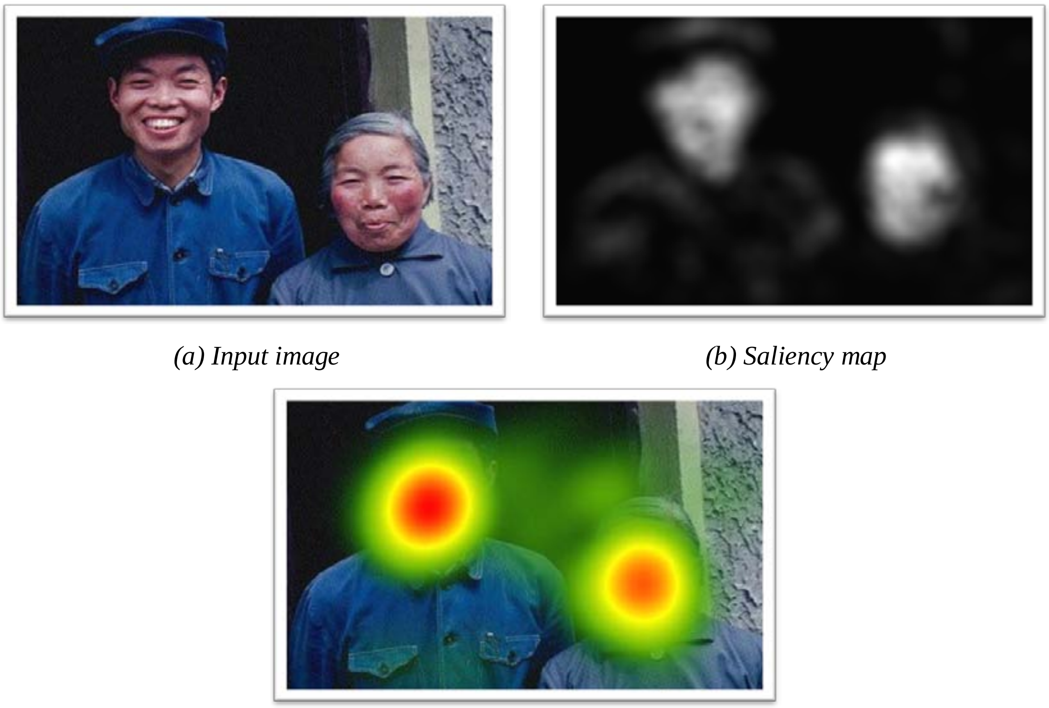
(a) Input image

(b) Saliency map

(c) Heat map

Figure 4-6 *Experimental results of natural image (IMG - 3)*

In Figure 4-7, which contains faces, the saliency map indeed pops these two faces out. The result is consistent with the human eye fixation result, which indicates that human faces would always be the visual saliency regions.

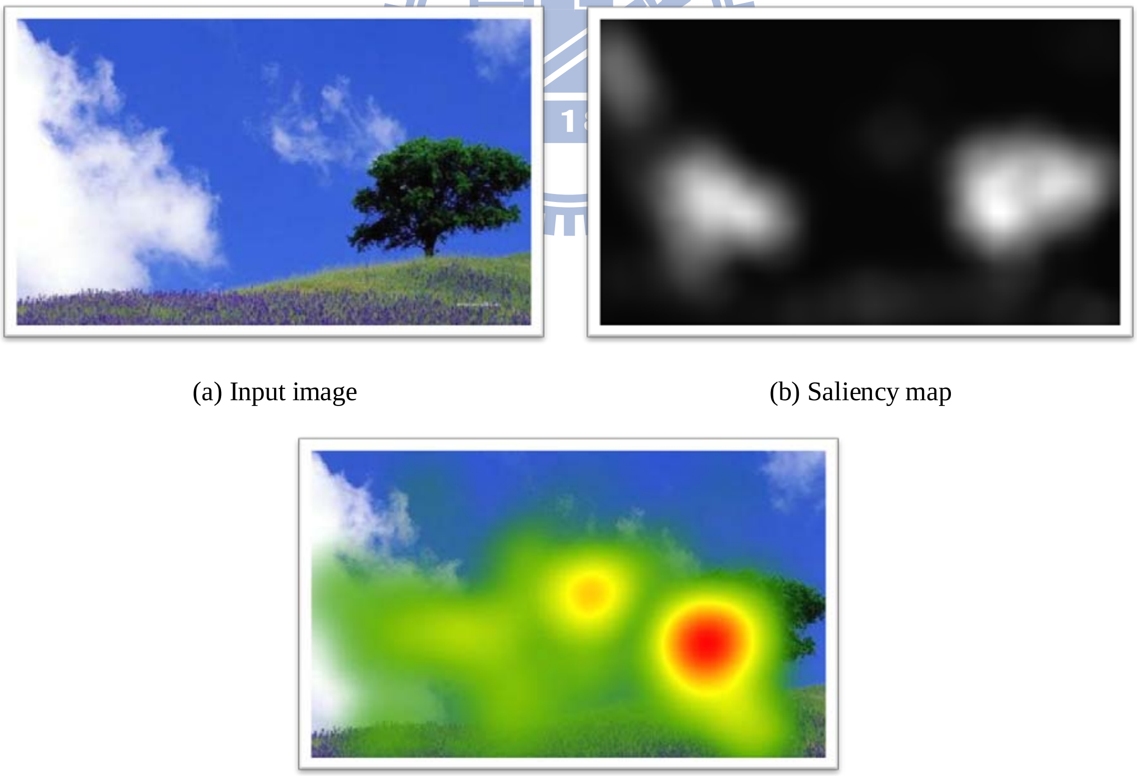


(a) Input image

(b) Saliency map

(c) Heat map

Figure 4-7 *Experimental results of image containing faces (IMG - 4)*

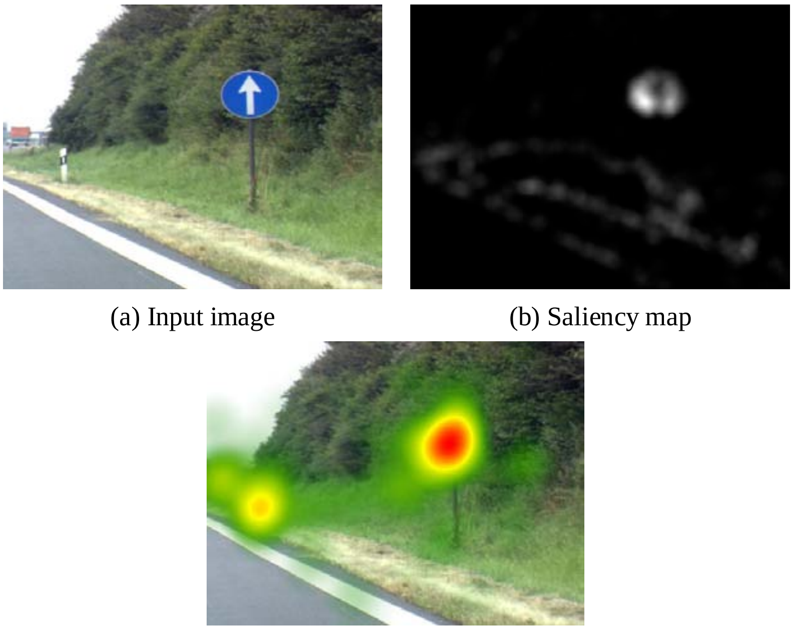


(a) Input image

(b) Saliency map

(c) Heat map

Figure 4-8 *Experimental results of natural image (IMG - 5)*



(a) Input image

(b) Saliency map

(c) Heat map

Figure 4-9 *Experimental results of natural image (IMG - 6)*

In Figure 4-10 to Figure 4-13, we show the performance comparison of our method with respect to the subjective experiment, Itti's method [4], and the Spectral Residual method [7], over four different images. The upper left image is the original image. The upper right image is the averaged eye fixation data, averaged from 20 subjects, with the red color indicating the visually salient regions. The detection results of the Itti's method, the SR method, and our method are shown in parallel for comparison. It can be seen that the proposed method outperforms both Itti's method and the SR method in these four cases. The results generated by Itti's method are somewhat different from the eye fixation data, while the results generated by the SR method are more like the results of edge detection. Moreover, the computation complexity of the proposed method is much lighter than that of Itti's method.

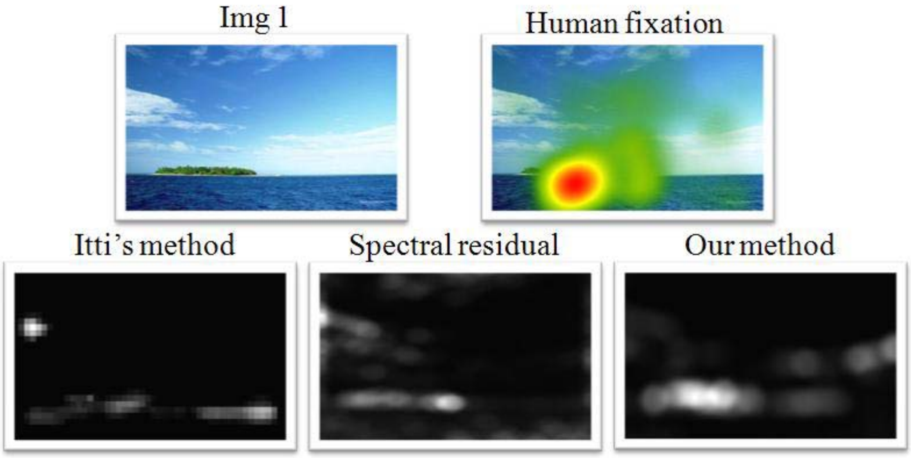


Figure 4-10 *Experimental results of comparisons with other methods (comparison – 1)*

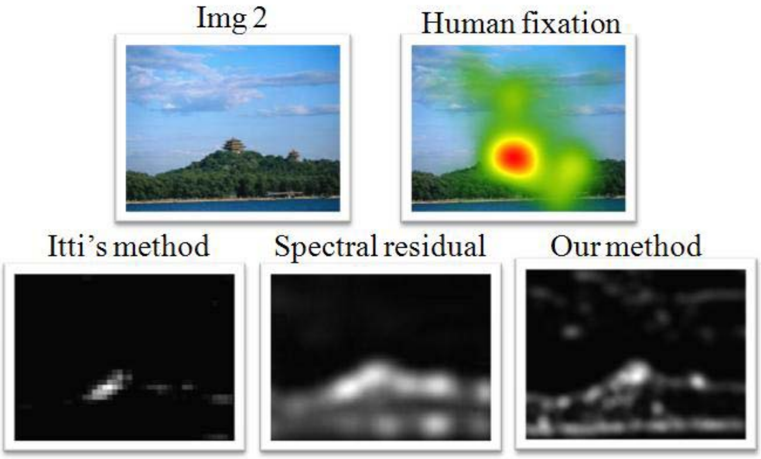


Figure 4-11 *Experimental results of comparisons with other methods (comparison – 2)*

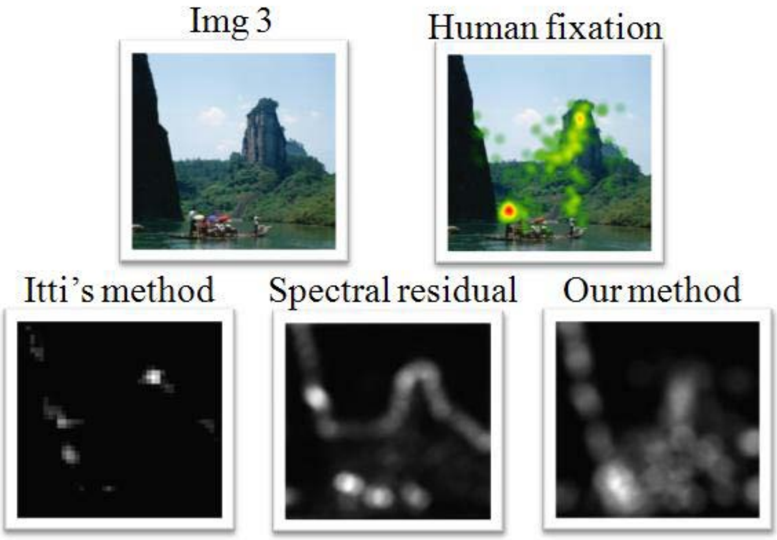


Figure 4-12 *Experimental results of comparisons with other methods (comparison – 3)*

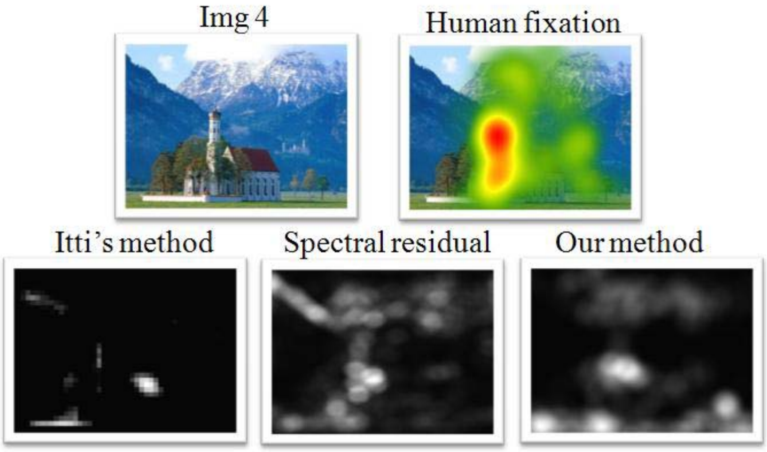


Figure 4-13 *Experimental results of comparisons with other methods (comparison – 4)*

Chapter 5.

CONCLUSIONS

In this thesis, we proposed a bottom-up feature-based technique for saliency region detection. The whole process is simple and doesn't require the training stage. For system activation, we extract the feature-pair distribution from low-level image data. We assign proper weights over the feature-pair distribution to identify visually salient regions. The proposed algorithm is much simpler than the commonly used Itti's method. After the activation process, a normalization process based on spatial competition is applied to the conspicuity maps to enhance signal-to-noise ratio. The conspicuity maps from different feature channels are then linearly combined in a data-driven manner. The experiment results show that the proposed algorithm can faithfully detect the salient regions for various kinds of images and the detection results are consistent with subjective observations.

REFERENCES

[1] E. Niebur and C. Koch, "Control of selective visual attention: Modeling the "where" pathway," *9th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 802-808, 1996.

[2] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency", *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

[3] C. Koch and S. Ullman, "Shifts in Selective Visual-Attention - Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.

[4] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.

[5] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Research*, vol. 40, pp. 1469-1487, 2000.

[6] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395-1407, Nov. 2006.

[7] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2280-2287, 2007.

[8] C. Guo, Q. Ma and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2908-2915, 2008.

[9] A.G. Leventhal, "The neural basis of visual function," *Vision and visual dysfunction*, Boca Raton, CRC Press, vol. 4, 1991.

[10] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," California Institute of Technology, Jan. 2000. [Ph.D. Thesis]

[11] L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," *Conference on Human Vision and Electronic Imaging IV*, pp. 473-482, 1999.

[12] T.C. Jen, B. Hsieh and S.J. Wang, "Image contrast enhancement based on intensity-pair distribution," *IEEE International Conference on Image Processing*, vol. 1, pp. I-913-16, Sep. 2005.

[13] S. Engel, X. M. Zhang and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, pp. 68-71, Jul. 1997.

[14] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.

[15] http://en.wikipedia.org/wiki/Eye_tracking