HIGH-DEFINITION HUMAN VISUAL ATTENTION MAPPING USING WAVELETS

by

Yusuf Saber, B.Eng.
Bachelor of Engineering (B.Eng.), Ryerson University, 2009

A thesis
presented to Ryerson University
in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering.

Toronto, Ontario, Canada, 2011

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Signature

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature

Borrowers' Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

High-Definition Human Visual Attention Mapping Using Wavelets

Yusuf Saber

Master of Applied Science, Electrical and Computer Engineering

Ryerson University, Toronto, Ontario, Canada, 2011

In this work, three novel approaches to detecting visual attention in images are presented. The idea behind detecting areas within images or video that naturally attract a viewer's attention is based on the concept of generating pre-attentive saliency maps. Saliency, in and of itself, relates to some measure of "conspicuity" in the visual field and is believed to be an important precursor for many tasks in computer vision. One of the proposed methods in this thesis detects salient regions, while the other two detect salient edges. The classical approach to saliency detection proposed by Itti is extended by introducing wavelets as a lossless resizing tool while maintaining the aspect of biological inspiration. In addition to this, the spectral residual method and the frequency tuned method are modified using wavelets to allow for salient edge detection. Tests show that the proposed methods yield results that are not only comparable to leading, cutting-edge methods, but also exceed them in terms of correct and complete object detection as well as noise reduction.

Acknowledgments

Of course, all praise and thanks go to God before anyone else. Without Him, nothing is possible.

After God, I must thank my parents Ramadan and Mervat Saber, and then the rest of my family, Donia, Amr, Mariam, Adam, Mallak, Ahmad, Chazia, Zackariya, and Mariam 2. They encouraged me to pursue my graduate studies, and continue to do so to this day. They helped with anything and everything I ever asked of them. Mama and Baba (and everyone else), thank you for everything. I hope I can one day repay you for all your help.

I must also thank my esteemed supervisor, Dr. Matthew Kyan. Without him, you would not even be reading these words! Matt, thank you for first giving me the opportunity to pursue Master's and then for allowing me to open my mind and really find myself through this work.

To all my friends and colleagues who I cannot mention, forgive me, there are too many of you. However, an honourable mention has to go out to Hussein, Mohannad, Aamir, Qureshi, Mansoor, Wazim, Osama, Azzam, Nabil, Emad, Shoaib, Boonaa, Ron-Ron, Ryan, Gopes, Ravi and this list won't end if I don't stop. I must also thank everyone in the multifaith and in the RAC, which is where I spent most of my time during my stay at Ryerson. And of course, I can't forget my partners during my undergraduate Engineering Design Project. Thank you Nazuk and James. To everyone in all the labs on the 4th floor (and EPH), thank you. Especially Raymond Phan, who I can honestly call my mentor and inspiration to continue into grad school.

These acknowledgements would not be complete if I did not mention and thank my fiancee, (soon to be Dr) Shaimaa. Although you only recently entered into my life, I know you will physically harm me if I don't mention you. So, thank you, and I love you from the bottom of my heart :)

"Everyone knows what attention is.
It is the taking possession by the mind,
in clear and vivid form, of one out of what
seem several simultaneously possible objects
or trains of thought. Focalization, concentration,
of consciousness are of its essence. It implies withdrawal
from some things in order to deal effectively with others, and is
a condition which has a real opposite in the confused, dazed, scatterbrained state."

-William James, 1890

Contents

List of Figures

List of Tables

Chapter 1

Introduction

ACCORDING to William James, the father of American psychology, a two component framework is implemented in the human visual system (HVS) for attentional deployment [7]. This framework suggests that the HVS directs attention to an object based on bottom-up cues as well as top-down cues [8]. Bottom-up cues (sometimes referred to as "pre-attentive" cues) are those that unintentionally grab one's attention, such as a bright area in an otherwise dark scene. Top-down cues are those that are intentionally looked for, such as trying to find someone in a room full of people.The top-down attention model thus requires prior information (knowledge of what is being searched for) in order to be implemented, while the bottom-up model does not. The focus of this thesis is bottom-up attention.

Certain situations always stimulate the HVS. For example: a bright red leather jacket amongst dull black ones, a small bright light in a dark room, or a car driving in one direction surrounded by other cars driving in the opposite direction [9]. To digitize the perception of attention, these stimuli must be modeled and captured as accurately as possible.

In the digital realm, saliency maps are used to determine possible regions of attention in an image. A saliency map is a 2D energy function which maximizes where the most attentive region exists. In terms of a human's gaze, the location on the saliency map that yields the highest energy is in essence, modeling where the gaze would unintentionally be directed.

Figure 1.1 shows an example of an image and its saliency map. The lighter the region is in the saliency map (the whiter it is), the more attentive the region is in the image. The

Figure 1.1: An image and its saliency map.

saliency map, which was produced using Itti's method (explained later), shows that the region of greatest attention is the white sign, followed by the phone, and finally a part of the road.

Although various methods exist for the extraction of salient objects, they all have the same fundamental backbone: to compare local regions with the global region. The difference between the local and global is directly correlated to the local region's level of saliency, where regions that are largely different than the global region yield a higher saliency. Some of the more basic methods try to extract salient regions by computing some representation of the overall image and subtracting each pixel in the image from that value [10]. Equation 1.1 gives this formula, where α is the representation of the entire image (could be the average pixel value of the image), S_p is the saliency of pixel p, and i_p is the intensity of pixel p.

$$S_p = i_p - \alpha \tag{1.1}$$

Other more advanced methods consider saliency in the context of spatial frequencies. For instance, in [5], sparsely-occurring frequencies are extracted from the image. In [4], the image is traversed and small local region are spatially compared to larger local regions; spatial comparisons may be performed with respect to various features of the image such as intensity, colour, orientation, hue, and/or texture. While the methods traversed in this thesis are the primarily fundamental ones, many more exist. Most of these are simple variations of others such as [11] which is based on maximum symmetric surround, [12] which is based on colour saliency (using wavelets), [13] which uses a fuzzy-growing algorithm alongside a contrast-based operator, [14] which utilizes images' properties of behaving like graphs, [15] which

also focuses on colour but uses a hill-climbing algorithm, [16] which is based on information maximization, [17] which uses integral images, [18] which focuses on finding salient objects solely in natural images, and [19] which focuses primarily on the idea of weighted feature maps.

1.1 Attention in the Human Visual System

As previously mentioned, visual attention is a two-component framework whose constituents are bottom-up attention and top-down attention [7]. When a stimulus is sufficiently salient, it will appear to "pop out" of a visual scene. This implies that saliency is computed pre-attentively across the entire visual field [1]. This type of visual attention is referred to as bottom-up attention. The visual system scans the entire visual field and attaches a level of saliency to everything, and then focuses on the most salient object.

The second component to the visual attention framework is top-down attention which requires work or higher level control from the human brain and is not done automatically. Whilst bottom-up attention seeks to find the most salient object regardless of what that particular object is, top-down attention seeks to find a particular object, regardless of its level of saliency. Both bottom-up and top-down saliency work in parallel in the HVS [1].

With regards to saliency itself, it differs with respect to top-down and bottom-up attention. With any given image, the bottom-up component will always make certain areas stand out, regardless of any pre-determined goal (such as finding something in the image). On the other hand, the top-down component will always try to suppress low-level features in the image in order to determine the location of some pre-determined object. The focus of this thesis is the development of algorithms that attempt to simulate the bottom-up component of visual attention.

1.1.1 Neuronal Mechanisms For The Control Of Attention

Since this is out of the scope of the thesis, but is an integral component to the visual attention system, a brief discussion is provided. It should be noted that this text is largely in reference

to [1].

Most of the early visual processing that occurs in the brain is shown in Figure 1.2. Through the primary visual cortex, outside information enters into the visual system. From there, visual information traverses through two main streams. Spatial localization of the gaze primarily comes from the cortical areas along the 'dorsal stream', which includes the posterior parietal cortex (PPC). As a result of this, the dorsal stream is believed to assume control of attentional deployment. On the other hand, recognition and identification of visual stimuli is controlled by the cortical areas along the 'ventral stream', which includes the inferotemporal cortex (IT).

As far as the general understanding of the computation of attention goes, it appears that the dorsal and ventral streams are required to interact in order for scenes to process. This is because scene understanding involves both recognition and spatial deployment of attention. An area where said interaction has been studied extensively is the prefrontal cortex (PFC). Within the PFC, areas are bidirectionally connected to the PPC as well as the IT. Thus, alongside being responsible for planning action (execution of eye movements through the SC, for instance), the PFC also has the important task of modulating, via feedback, the dorsal and ventral processing streams.

1.2 Applications

There are many applications that use saliency maps such as unsupervised image segmentation [20] [21] [15], robotic vision [22], image compression [3], surveillance [23], and object detection [24].

A major area of application is the mapping of eye movements using heat maps (see Figure 1.3). Heat maps show the locations where a human's gaze is directed [2]. The eye movements can be recorded using a number of methods, including a simple recording using a video camera. Rather than recording eye movements to yield a heat map, saliency maps could predict where eyes will be directed within a scene by locating the attentive regions.

Based on the specific application, different types of saliency maps could be applied. For

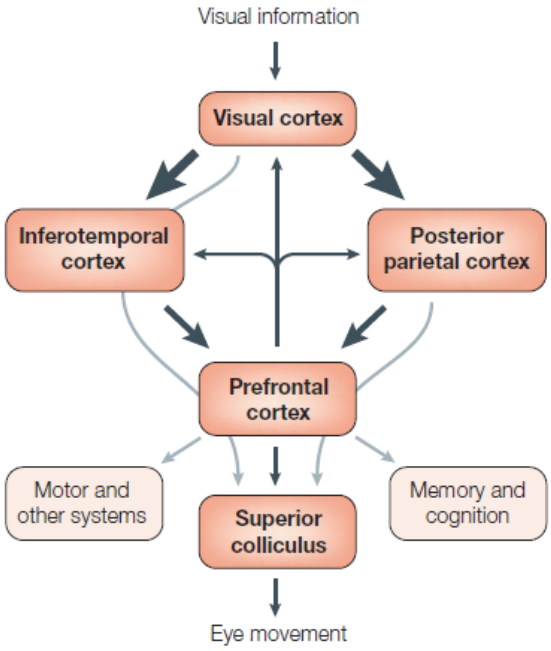


Figure 1.2: Neuronal mechanisms for the control of attention [1].

example, with regards to robotic vision, low-resolution maps suffice. With regards to image compression or segmentation, high-resolution maps are more appropriate. Visual attention cannot be limited to any limited number of applications as it can be applied to the digitization of anything that requires a model of the HVS, and plenty more. Select examples of some applications are elaborated on below.

1.2.1 Robotic Vision

In [22], the authors used off-the-shelf parts to build a small, light, powerful, and inexpensive robot that uses visual attention to direct it; they called it the *Beobot*. It is titled as such because it is based on their lab's work on a system called *Beowulf* which links multiple computing units into a single cooperative system. Their goal for this project was to demonstrate a new robotics platform with sufficient computing resources to run biologically-inspired vision algorithms in real-time, and they accomplished just that. Using a Unibrain Fire-I camera, the scene ahead of the robot is captured and the following algorithm is applied:

Figure 1.3: Heat map [2].

1. Scan the image and do an initial thresholding on pixels based upon some desired threshold.

2. Remove candidate pixels without at least one 4 connectivity neighbor since it is most likely noise.

3. Scan the image and link candidate pixels into discrete blobs. Link connected blobs into unified blobs.

4. Based upon mass, dimensions, ratios and trajectories with hard constraints or soft statistical constraints, remove blobs less likely to be the target.

The robotic car was tested in outdoor navigation and exploration activities such as follow-the-leader type tasks.

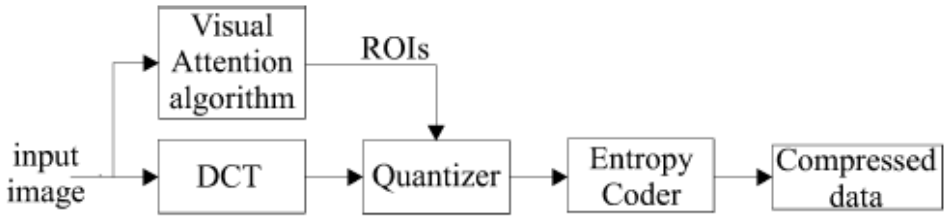


Figure 1.4: Adaptive JPEG-based image compression algorithm [3].

1.2.2 Image Compression

In [3], the authors use visual attention (Itti's model) to locate regions in an image that are more prone to viewer-interest. Using this information, they apply JPEG compression to the entire image, with less emphasis on the locations of attention (referred to as regions of interest (ROIs)). This results in a smaller compression ratio within the locations where visually attentive objects exist, allowing for a more accurate result when the image is reconstructed.

Figure 1.4 shows a block diagram of the adaptive coding method. This scheme follows the same operations of the baseline JPEG algorithm, albeit with a quantization unit that has been modified to receive an additional input: the saliency map produced by the visual attention block. The saliency map tells the quantizer to execute either a short-step or a large-step quantization of the DCT coefficients, depending on whether a given 8x8-pixel block lies inside or outside any of the identified ROIs.

1.2.3 Unsupervised Image Segmentation

In [20], the authors approach unsupervised image segmentation using a two stage framework. First, they use saliency to locate the attentive regions. Second, they use fuzzy support vector machines (FSVM) to outlines (segment) the object from the remainder of the image.

The authors use Itti's approach to finding salient regions, but they take it a step further by searching for corner points in order to locate a rectangular region of attention (ROA). This prepares the image for the FSVM, which uses both the image inside the ROA as well as pixels outside of it to extract a foreground object from the background within the ROA.

1.2.4 Surveillance

In [23], the authors use motion saliency to locate peculiar objects in a sequence of video. A motion saliency map is simply a map that displays salient motion within a few frames of a video. The authors use a difference operator across all the pixels to determine saliency in the pixels. The difference between an attentional approach and standard motion segmentation being that the former only considers unusual motion.

Since surveillance targets are often the single object that is moving drastically in a video, this approach is sufficient for its purpose. Some post processing is implemented on the difference map to make the moving object appear more vividly, but that is beyond the scope of this thesis.

1.3 Thesis Overview

The objective of this thesis is to first comparatively discuss a few prominent saliency extraction methods from the literature, then describe some novel advancements, and finally compare results of the proposed methods with those from the literature.

Chapter 1 (Introduction) gives a high-level explanation of what visual attention is, and what saliency is. Explaining the attention component in the human visual system, this chapter attempts to solidify the basis of the models that are explained in forthcoming chapters, namely Chapter 2 (Past Approaches), and Chapter 3 (Novel Advancements). The transition from scientific understanding to computational model is also described. Following this, some applications are discussed, to further cement the ideology of what will be discussed in later chapters.

Chapter 2 (Past Approaches) discusses three of the most prominent approaches taken towards discovering saliency in images, organized chronologically. The first approach discussed is proposed by Laurent Itti, Christof Koch, and Ernst Niebur in 1998 [4]. This approach is based entirely on the physiological understanding of the HVS. Features (intensity, colour) are extracted using linear filters, and the centre-surround operation is applied to each feature

separately. This results in a series of conspicuity maps (saliency map of a single feature), which are combined to yield a final saliency map for the image. The second approach is motivated by computational efficiency and was proposed by Xiaodi Hou and Liqing Zhang in 2007 [5]. This approach attempts to extract sparsely-occurring frequencies from a down-sampled version of the image. By doing this, the object or region that occupies the most oft-occurring frequencies is extracted. The final approach that is discussed in this chapter is based on producing a high-resolution, computationally-efficient saliency map. This method was proposed by Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk in 2009 [6]. The authors proposed a set of rules that their saliency maps should abide by, and then proposed a method to follow the rules. They apply a low-pass Gaussian filter to slightly blur the image (in order to remove noise), and then apply a frequency-tuning operation which, in a way, is a correlation function. Each section in this chapter also includes a few sample results.

Chapter 3 (Novel Approaches) traverses the author's original contributions to the literature of visual attention. There are three proposed methods: the High Definition Human Visual System (HDHVS) method, the Wavelet Residual (WR) method, and the Frequency-Tuned Wavelet Residual (FTWR) method. The HDHVS method detects salient regions in an image while the WR and the FTWR detect salient edges. The idea of salient edges has never been discussed in the literature before. The HDHVS method uses Itti's method as a basis, but uses the discrete wavelet transform (DWT) as a lossless resizing tool. The WR uses the SR as a basis, but, again, uses the DWT as a lossless resizing tool. Finally, the FTWR has bases from both SR and Achanta's method, and again, uses wavelets as a lossless resizing tool.

Chapter 4 (Results) shows a multitude of results from each approach/method for the reader to compare, along with a discussion of each of the results. A discussion on the evaluation of the results is also presented.

Chapter 5 (Conclusions and Future Work) discusses road blocks and limitations of the methods discussed in this thesis, and potential measures that can be taken to enhance them.

Colour opponency is paid particular attention to in this section, but other topics such as time-frequency conversion is also discussed.

Chapter 2

Past Approaches

HISTORY tends to repeat itself. If we reinvent the wheel, we have done nothing but delay the inevitable progression of technology. Studying the history of something allows for a thorough understanding of it's principles, purposes, and results; all of which are key elements to the thing's progression. Visual attention is no stranger to this scheme of progression. As we will see in this section, Itti et al used physiological concepts of visual attention to model the first electronic version of the human attention system [4]. From there, the concept of centre-surround operator was born. Computational efficiency then became an issue of concern and so the spectral-residual method was invented by Hou et al [5]. Once the significance of saliency maps became apparent, higher quality maps were necessary, and so Achanta et al defined a set of rules that the maps must abide by, after which he invented the frequency-tuned method [6]. In this thesis, novel upgrades are suggested which may be the next steps in the evolution of the progression of visual attention in images.

This chapter begins with a thorough walk through of Itti's human-visual-system-based method, since it is the foundation of all other methods. In this discussion, an overall glimpse of the method is employed and then details of each component are traversed. From there, the transition to the spectral-residual method is shown. Soon after which the progression to the frequency-tuned method is discussed. At the end of each section, sample results are shown, so that the reader can have a better appreciation for the method at hand.

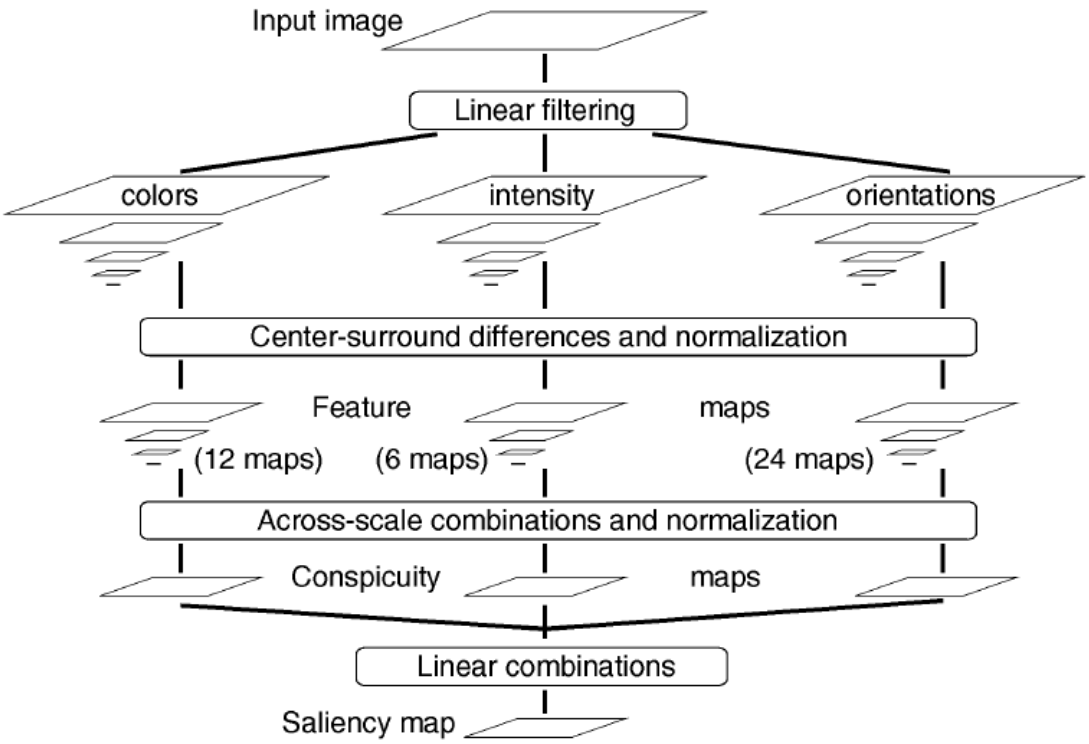


Figure 2.1: Block diagram of Itti's method [4].

2.1 Laurent Itti's Method

To introduce this method, a block diagram of the entire process is presented in Figure 2.1. This block diagram was designed by Itti in his literature [4]. It should be noted that in the original literature, this figure contains more components, but for the purpose of this thesis, only the relevant components for bottom-up processes have been included.

From the input image, multiple linear filters are applied to extract the selected feature information (in this case: intensity, colour, and orientation). The extracted information is recorded in a two-dimensional image, or map. To extract salient points from each of the features, the centre-surround operation is implemented on each of the feature maps, at multiple scales, for true multi-scale comprehension. The result of the centre-surround operation is a feature map which is also a 2D image/map containing salient points for a specific feature. At the end of the centre-surround computations, there will result 6 feature

Figure 2.2: Image pyramids showing scales 1(original), 2, 3, 4.

maps for intensity, 12 for colour, and 24 for orientation. Combining the feature maps yields 3 conspicuity maps which show the conspicuous regions with respect to each feature. These images are then normalized and combined to produce the final saliency map.

2.1.1 Centre-Surround Operator

The centre-surround mechanism used by Itti can be defined as the difference between a small centre region and its close surrounding. This is based on the fact that differences in either intensity or some other property at different scales are known to trigger neural responses in the human visual system. It is implemented using Gaussian image pyramids as a mechanism for creating various scales of the image. These pyramids are created by down sampling the image. For our purpose, the image is down sampled eight times, each time by a factor or two. Figure 2.2 shows an image of a bird and three levels of its pyramids.

The reason image pyramids are used is to simulate the effect of the surrounding more

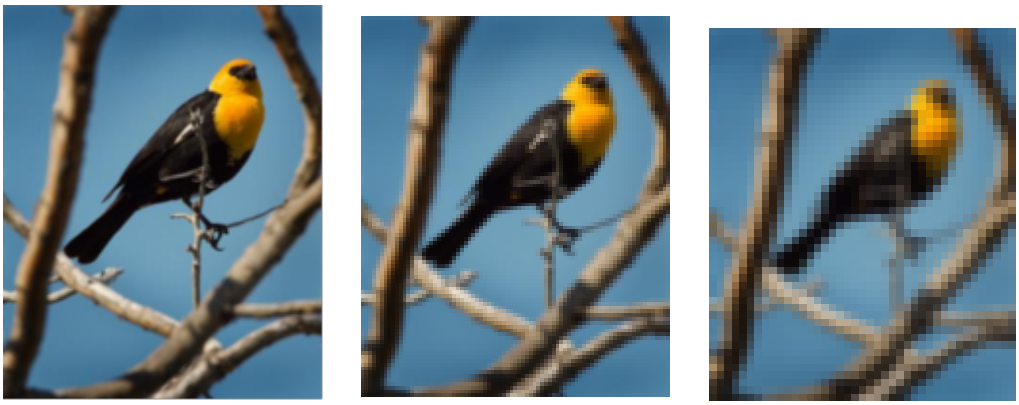


Figure 2.3: Image pyramids resized.

accurately. Intuitively, one might assume that one can just subtract the sum of the pixels in a small region from the sum of pixels in a slightly larger region. Although this works, it is more desirable to consider a region in and of itself. This may seem like an oxymoron, however when the image pyramids are created, what is actually being created is a blurred representation of the original image, which simulates a local-surrounding operation since each pixel in the image is replaced by the average of its neighboring pixels when downsampled.

Figure 2.3 shows the second level of the pyramid in Figure 2.2 as well as the preceding two levels, all with the dimensions of the second-level image. Traversing these images (called *scales* of the original image) two at a time and performing pixel-by-pixel subtraction is the essence of the centre-surround function.

The centre-surround function is defined as the difference between some property of the image at a finer and a coarser scale of the original image, with respect to a particular feature or property. Let's take the pixel intensity or colour as the property, for instance. The scale at which the centre is taken is the finer scale while the scale at which the surround is taken is the coarser scale. The centre is a pixel at scale C, and the surround is the corresponding pixel at scale S, where $S = C + \delta$ where $C = 2, 3, 4$ and $\delta = 3, 4$. The across-scale difference between two scales is obtained by interpolating the coarse scale to the finer scale and performing point-by-point subtraction.

Using several scales yields multi-scale feature extraction by including different size ratios

between the centre and the surround regions. It should now be clearer why image pyramids are used rather than subtracting the sum of the pixels in a small region from the sum of pixels in a slightly larger region, because the former compares a representation of a given region to the region itself, while the latter compares two entirely different regions.

This process of finding the centre-surround difference must be applied to a certain property or feature. As aforementioned; intensity, colour, and orientation are the features that are used in this work. They are discussed in the following section.

2.1.2 Local (Centre) vs. Global (Surround) Feature Selection

Itti's method is built, in part, based on the classical Feature Integration Theory which was developed by Treisman in 1980 [25]. It states that visual input is first decomposed into a set of topographic feature maps. Therefore, to correctly model the visual system, the centre-surround operator must be applied to features individually, and then combined to form a master saliency map.

Of the many features that can be used (hue, texture, etc.), intensity, colour, and orientation appear to provide the most accurate results. Because of this, they are the only features that are considered in the implementation in this thesis. Other features have been used however, and are available for use online (see [26]).

Intensity Contrast

In the literature presented by Itti, it is suggested that intensity be computed by taking the average of the three colour channels, as is shown in Equation 2.1, which results in a grayscale map of the input image.

$$I = \frac{(r + g + b)}{3} \tag{2.1}$$

While the model in Equation 2.1 might be intuitive, Equation 2.2 provides a more accurate array of the luminance in the image.

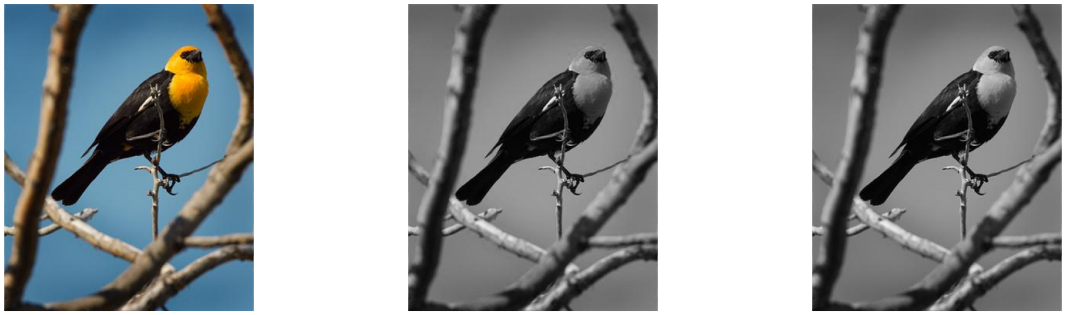


Figure 2.4: Comparison of RGB-to-grayscale functions. Middle image is using Equation 2.1. Right image is using Equation 2.2.

$$I = 0.298r + 0.578g + 0.114b \tag{2.2}$$

These weights are used in the conversion from the RGB colour space to the YUV colour space, where I in the above equation corresponds to the luminance (Y-component) in the YUV space.

Equation 2.2 is more accurate than Equation 2.1 because of the fact that the channels in the RGB space are not equal constituents of the colours they combine to make. Rather, the red channel represents 29.89%, the green represents 58.7%, and the blue channel represents the remaining 11.4% of the desired colour.

Colour Opponency

The second feature that is considered is colour. Colour is a complex phenomenon that is related to the fusion of some responses in the retina. Colour is something that is "'perceived"' in the brain and is thus due to cognition, not simply stimulus alone. Although the HVS is attracted to certain colours over others (red, for example), the more impacting aspect of colour is that of certain opponencies [27]. A colour opponency is basically a colour contrast (or dissimilarity) between two spatially close regions. There are two main opponencies are discussed in the literature [4], they are: blue-yellow and red-green. This means that when a blue region is spatially close to a yellow region, the HVS is strongly attracted. The same

goes for red and green regions that are spatially close. This concept of colour-opponency is the foundation of extracting the colour feature from an image.

Since one of the opponencies requires the yellow information of the image to be extracted, a process must be brought forth to handle such a task. The RGB space, as the name indicates, directly represents only red, green, and blue channels, so a manipulation must be implemented to observe yellow information. These channels are not particularly true representations of their respective colours either. The reason they are split as such is because using these three colours, virtually any colour can be created. Referring back to the discussion concerning the conversion from RGB to grayscale, each of the channels in the RGB space is not an equal constituent of the colours they combine to make. Hence, thinking the red channel contains information of only reddish pixels, the green channel contains only greenish pixels, and the blue channel contains only blueish pixels is quite naive. These channels only comprise of the constituent that is required of it to formulate a given colour since colour is a continuum (a perceived similarity as the electromagnetic spectrum is traversed). Colour opponency in some sense relates to the relationship between certain wavelengths of light and how they exhibit a strong dissimilarity in terms of perceived appearance with respect to one another (such as violet vs. yellow).

Since our goal is to extract true red, green, and blue information, some conversion is required. In his literature, Itti proposes Equations 2.3 - 2.6 to extract true red (Equation 2.3), green (Equation 2.4), and blue (Equation 2.5) colours from an image. In these equations, the lower case letters refer to the channels in the RGB space while the upper case letters refer to the new colour representation image/map. Equation 2.6 is the function proposed to extract yellow colours from an image. This is required to implement the blue-yellow colour opponency component.

$$R = r - \frac{(b+g)}{2} \tag{2.3}$$

$$G = g - \frac{(b+r)}{2} \tag{2.4}$$

$$B = b - \frac{(r+g)}{2} \tag{2.5}$$

$$Y = \frac{(r+g)}{2} - \left|\frac{r-g}{2}\right| - b \tag{2.6}$$

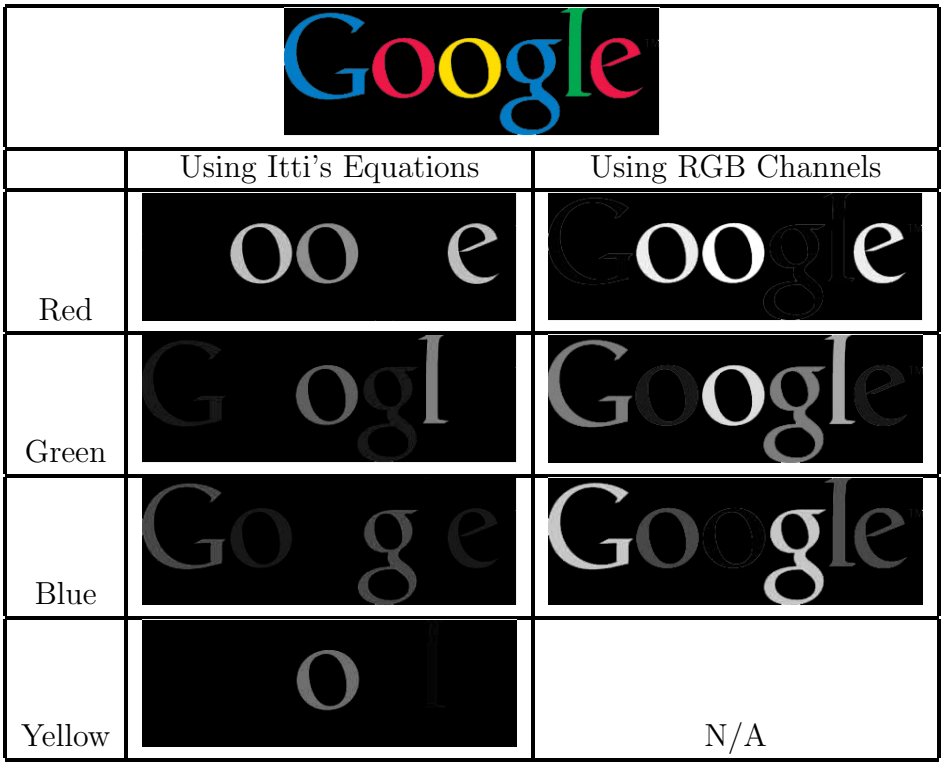
Although no derivation is provided for these equations, they can be verified using a test image. This is demonstrated in Table 2.1. Although Equations 2.3 - 2.6 extract colour information more accurately than using the raw image channels, it should be noted that they are not perfect. This can be seen in the red and green rows in Table 2.1, where, the images in Itti's column contain yellow colours when only red and green colours, respectively, should be visible. This happens due to the fact that there are many types and levels of red, green, and blue.

Orientation Opponency

The final step in extracting the colour opponencies is done during the centre-surround function. The opponencies are implemented by having one colour feature map at one scale and the other colour feature map at the other scale. For example, to implement the blue-yellow opponency for centre scale 4 and surround scale 7, the 4th scale of the blue image pyramid would be used alongside the 7th scale of the yellow image pyramid and that would be followed by the regular interpolation and point-by-point subtraction.

The final feature to be considered is orientation. As mentioned in the introduction, when objects are oriented different than the norm in a scene, they stand out to the HVS. Figure 2.5 shows an image with arrows, the majority of which are pointing up. Two of those arrows however are pointing to the right and they stand out due to their peculiar orientation. This can be concluded to be the feature of distinction since all the arrows are otherwise identical, having the same colour, size, and background.

The orientation computations are performed using the intensity image I since orientation does not change with colour. Orientation is extracted from an image using an edge detection function called the Gabor filter [28]. By performing the centre-surround operator on various



	Using Itti's Equations	Using RGB Channels
Red		
Green		
Blue		
Yellow		N/A

Table 2.1: Itti's equations for colour extraction vs. raw colour channels

orientations of the image, attentive regions in the image with respect to their orientation are extracted.

The Gabor filter, named after Dennis Gabor [28], is a linear filter used for edge detection. In the spatial domain, a 2D Gabor filter is a scale dependant Gaussian kernel function that is modulated by a sine plane wave. Figure 2.6 shows four spatial domain representations of the Gabor filters with varying orientations (0°, 45°, 90°, and 135°). To detect edges that are at, say, 45° in an image, the second Gabor filter from the left in Figure 2.6 would traverse the entire image as any spatial filter would and convolve with each pixel. Alternatively (and this is done in practise for its computational efficiency), both the filter and the image can be converted to the frequency domain using a Fourier transform and then simply multiplied together. The result is an edge map containing edges oriented at 45°.

Creating the Gabor filter requires a number of variables to be tuned including the wave-

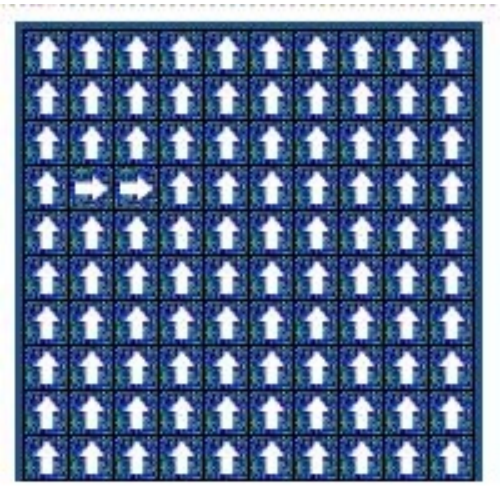


Figure 2.5: Orientation demonstration.

Figure 2.6: Gabor filters at orientations 0°, 45°, 90°, and 135°.

length of the sinusoid (λ), the orientation of the filter (θ), the phase offset (ψ), and the spatial aspect ratio (γ). Choosing these variables and plugging them into Equations 2.7 and 2.8 yields a Gabor filter that can then be convolved with an image to produce an edge map. Equation 2.7 is the real component of the filter while Equation 2.8 is the imaginary component. These two components can be combined to form a complex number or can alternatively be used individually.

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp\left(-\frac{x^{'2} + \lambda^2 y^2}{2\sigma^2}\right) cos\left(2\pi\frac{x^{'}}{\lambda} + \psi\right) \tag{2.7}$$

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp\left(-\frac{x^{'2} + \lambda^2 y^2}{2\sigma^2}\right) sin\left(2\pi\frac{x^{'}}{\lambda} + \psi\right) \tag{2.8}$$

where

$$x^{'} = xcos\theta + ysin\theta \tag{2.9}$$

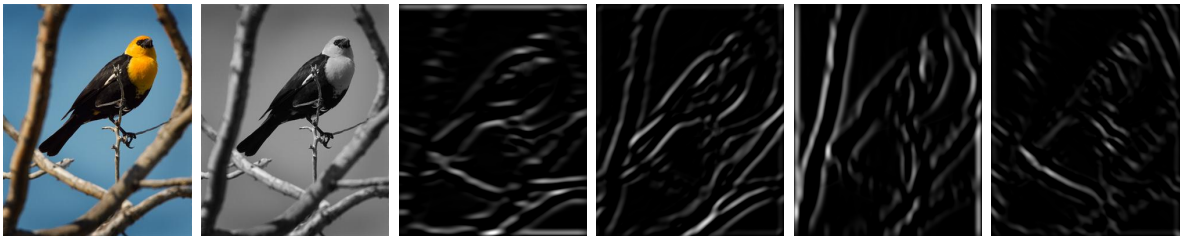


Figure 2.7: Yellow bird image and its edge maps at orientations 0°, 45°, 90°, and 135°.

$$y^{'} = ycos\theta - xsin\theta \tag{2.10}$$

Using the four mentioned orientations is sufficient to extract attentive regions with respect to their orientation, as suggested by Itti [4]. Figure 2.7 shows an image of a yellow bird and its respective edge maps at orientations 0°, 45°, 90°, and 135°.

It should be kept in mind that the Gabor filter is implemented in the frequency domain by means of Fourier transformation and multiplication with the Fourier transform of the image being investigated. This is done purely for computational efficiency.

The final step is to perform the centre-surround operation on each of the orientations. This allows for an identification of when a particular feature differs locally from the global norm. It is important to note that when doing this, the edge maps are not simply resized, rather a smaller Gabor kernel is produced and applied to a lower scale of the image. This allows for edges in that particular scale to be correctly extracted.

2.1.3 Normalization Operator

An important point to note is that some feature maps may yield varying dynamic ranges and a region that is salient but of a low level in one map may be masked by another map that is much stronger but has no salient points. Itti proposed a normalization operator in his literature which globally promotes maps in which a small number of salient regions are present while suppressing maps which contain numerous but comparable salient regions. If a map contains numerous comparable salient regions, it can be assumed that in an isolated space, each of these regions would be strongly salient but grouped with other salient regions,

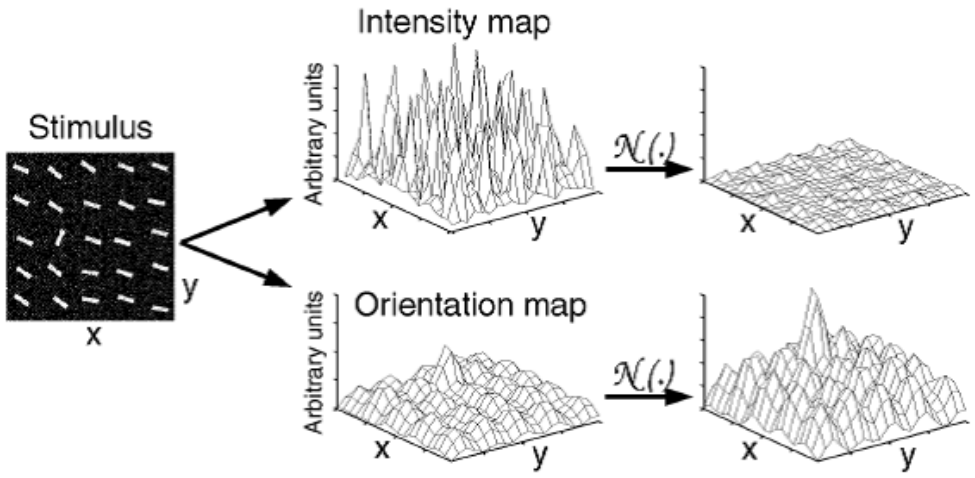


Figure 2.8: Itti's normalization operator [4].

attention is not drawn to one particular region, so the saliency is removed from the scene. For example, if Figure 2.5 contained a larger number of arrows pointing to the right, or if all the arrows were pointing in different directions, none would ultimately stand out.

Figure 2.8 shows an image titled 'Stimulus' which contains a noticeable region with respect to its orientation. In terms of intensity, the region does not stand out. However, computationally, each of the little strips induces a strong salient point due to its strong distinction from the background, but none stand out individually. The orientation map however does not induce a very strong response overall but the region where the odd strip is induces a stronger response than the rest because of its conspicuity (i.e. locally it is not usual with respect to the rest of the image). Because of this, the normalization operator is required in order to suppress the strong but comparable responses in the intensity map while promoting the small but unique spike in the orientation map.

The normalization operator is implemented in three steps:

1. Find the global maximum of the map, M.

2. Compute the average of a sample of the local maxima, m.

3. Globally multiply the image by $(M - m)^2$.

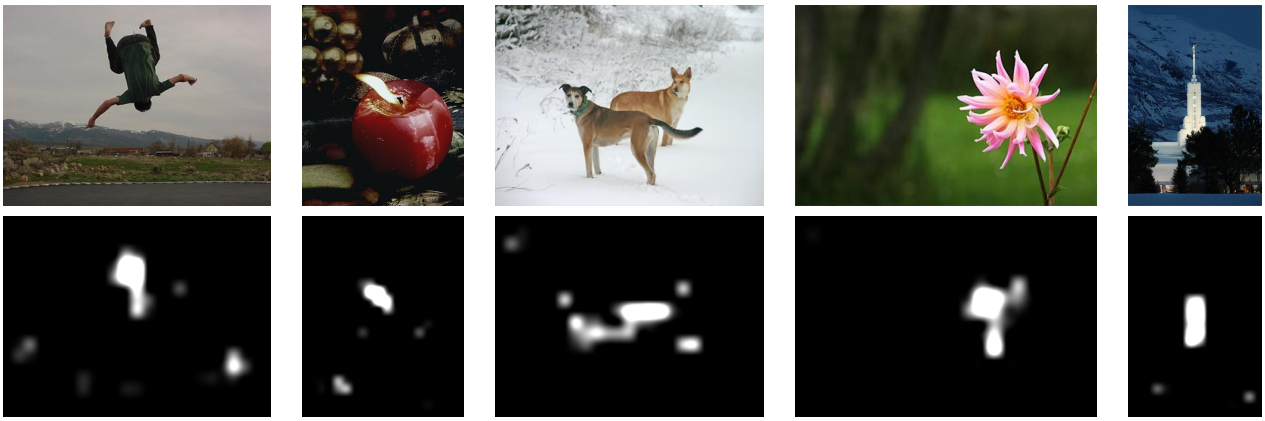


Table 2.2: Samples of Itti's method.

If the entire image is homogenous, $(M-m)^2$ will be very low and will hence suppress the map entirely. However, if $(M-m)^2$ is large, the conspicuous region (where M is located) will increase, since it also has a large value, while the rest of the image will be suppressed, since it has a low value.

2.1.4 Sample Results

Some results are shown to give a better understanding of Itti's method. The results are resized to the resolution of the input image using bicubic interpolation.

It is evident that Itti's method detects the location of the most attentive object in the image, but that is as far as it goes. The shape that represents the object is often not a good representation of the actual shape of the object, since such a large number of resizing operators are applied. This is mostly evident in the picture of the upside-down person, the two dogs, and the flower. It is important to note that Itti's method was originally used to correlate with, and model, gaze direction and saccadic eye movements.

2.2 The Spectral Residual Method

As mentioned in the introduction, the ultimate goal of a saliency map is to suggest regions that are both local and different from the norm of the global image. The fundamental prin-

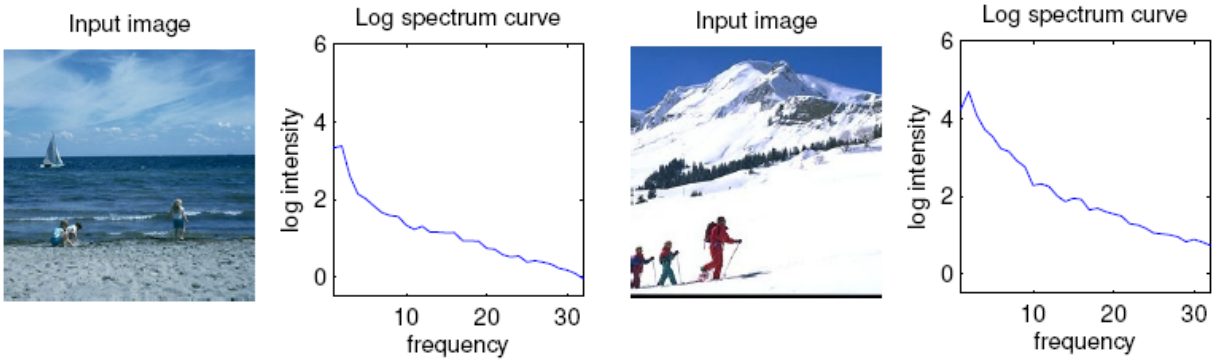


Figure 2.9: Log spectrums of natural images [5].

ciples behind which this is implemented will differ from method to method. This underlying goal is also true for the spectral residual method, which aims to locate frequencies in the image that differ from the global norm.

The spectral residual method [5] is based on the fact that the log-spectrum of any given natural image has an expected shape, of course with some minor differences in each image, as is shown in Figure 2.9. Notice that as the frequency increases, the number of occurrences of that specific frequency decreases. This phenomenon is a common trend among natural images which can be used to locate frequencies that are different from this trend. Removing the expected spectrum from the actual spectrum leaves behind a residue that is used to construct the saliency map. The spectrums shown in Figure 2.9 are represented as a 1D graph, however this reflects the profile of the 2D spatial frequency.

The expected spectrum can be produced by either taking the average spectrum of a large set of images, or by smoothing the image being investigated. This is shown in Figure 2.10. Due to its simplicity, implementing this method is done by smoothing the spectrum.

As can be seen in Figure 2.11, subtracting the log-spectrum of the image from the expected spectrum yields the aforementioned spectral-residual. Note that the plot titled "Spectral average curve" is the log-spectrum of the input image, but smoothed, not the log-spectrum of a combination of a large number of images.the latter is a more realistic implementation, however, is not as practical as the former, since a large number of images is

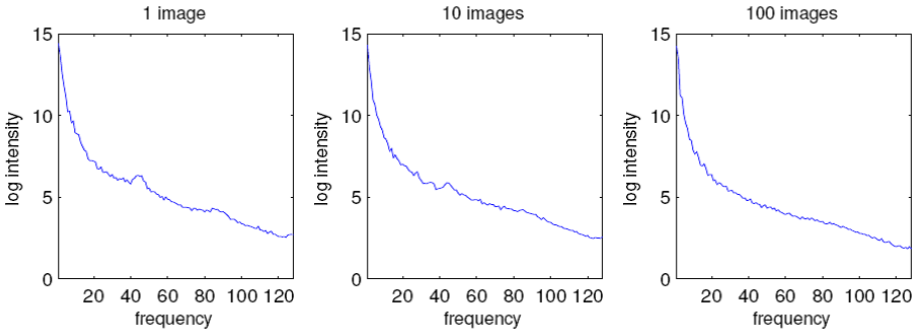


Figure 2.10: Similarity in log spectrums of many images [5].

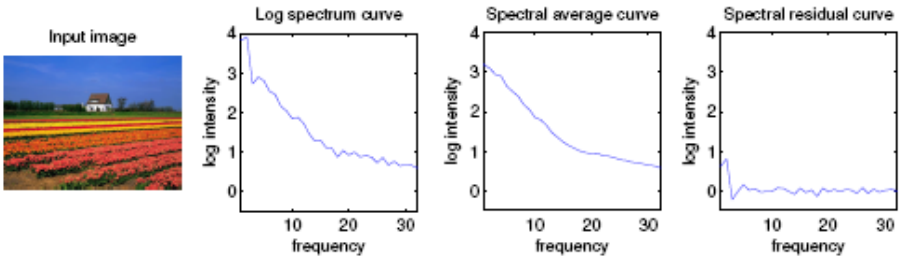


Figure 2.11: Extracting the spectral residual curve [5].

not always available.

Reconstructing the original image using the spectral residual as the amplitude rather than the original log-spectrum of the image yields the desired saliency map. An important point to note about this method is that it picks up uncommon frequencies, not uncommon regions. Because of this, the saliency map produced reflects, to some degree, scale based discontinuities that are infrequent in the image. These present themselves as an outline of the region of attention, rather than a blob like the one produced in the HVS method.

The entire process is demonstrated in Figure 2.12. This entire process takes place in the Fourier domain since that is where frequencies can be analyzed more readily. As a preprocessing stage, the image is resized to dimensions of roughly 64×64. This is done to remove small artifacts and noise. The second step is to compute the Fourier spectrum of the image using the 2D Fourier transform. Then the amplitude and phase are separated because the phase information is required to correctly reconstruct the image, so it is left

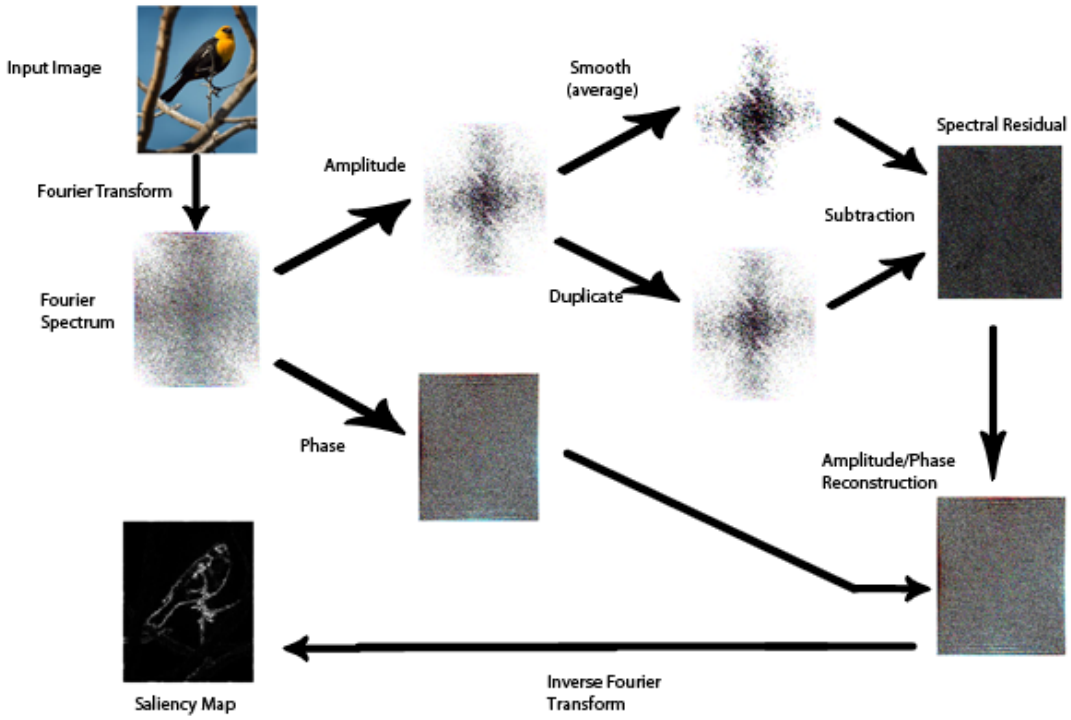


Figure 2.12: Block diagram of spectral residual method.

unchanged throughout the entire process. The amplitude is then smoothed, which produces the simulated expected amplitude spectrum. Subtracting this spectrum from the actual amplitude spectrum leaves the spectral residual which is then used alongside the original phase spectrum to reconstruct the image. When the image is reconstructed with the spectral residual rather than the original amplitude spectrum, most of the image's amplitude levels are very low (black pixels) except regions where the residual is strong, those regions have a higher amplitude (white pixels) and thus is the saliency map yielded. Pseudo-code for this algorithm is as follows:

1. $FFTimg = F(img)$
2. $A = Real(FFTimg)$
3. $P = Imaginary(FFTimg)$

4. $L = log(A)$

5. $R = L - (h * L)$

6. $S = g * [exp(R + P)]^2$

Comparing this method to the Itti's method, it is clear that the fundamental principles are different. Itti's method tries to mimic the HVS while the spectral residual looks at unusually occurring frequencies. Another important difference is that Itti's method yields blobs of salient regions while the spectral residual method yields outlines of salient regions.

Spectral Residual Optimization

It was discovered by Guo et al. [29] that rather than performing the SR method, the same result can be achieved by reconstructing the Fourier decomposition using only the phase spectrum. This means that the magnitude spectrum is not even considered. The algorithm is as follows:

1. $FFTimg = F(img)$

2. $P = Imaginary(FFTimg)$

3. $S = g * exp(P)^2$

The result of this optimization is surprisingly comparable to that of the original SR.

Modification to Extract Salient Edges

The concept of salient edge detection has not been discussed before, to the best knowledge of the author of this thesis. However, it is an important topic to research as detecting edges usually requires less processing time than detecting entire regions. Also, edge maps have many applications that salient region maps simply cannot satisfy.

Although no literature exists on this matter, a slight modification to the SR yields a salient edge map. It was discovered by the author of this thesis that applying the spectral residual method to an image without resizing it yields an edge map of the salient object.

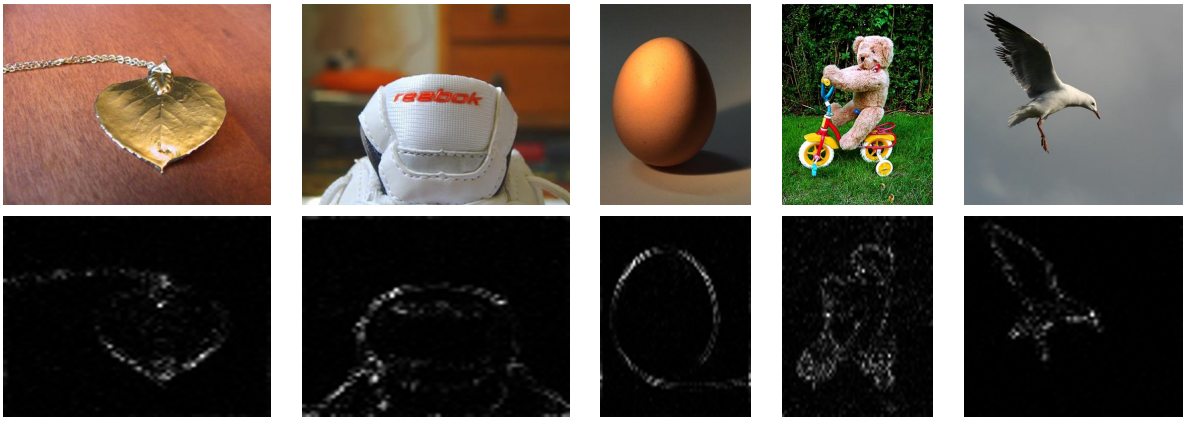


Table 2.3: Samples of the spectral residual method.

2.2.1 Sample Results

Some results are shown to give a better understanding of the spectral residual method. Again, the results have been resized to the dimensions of the original image using bicubic interpolation.

It can be seen how the SR has an outlining effect as a result of its extracting peculiar frequencies. Since the regions within these objects is much smoother than the outline, the regions are not picked up. Due to the resizing that the authors suggest, the saliency maps are quite blurry. It should also be noted that the SR method does a very good job at suppressing noisy object and artifacts in the images.

2.3 The Frequency Tuned Method

Achanta et al. [6] proposed a set of definitions to yield a more versatile saliency map. Their vision was to create a saliency map that is the same resolution as the original image, so that an exact outline of the salient region is extracted, rather than just a blob of the region where the salient object exists, which is the case with previous methods. Namely, Itti's method produces saliency maps that are just $1/256^{th}$ the original image size in pixels, while the Spectral Residual approach outputs maps of size 64×64 pixels, regardless of the input

image size. The authors put forth a set of requirements that the high resolution saliency map should adhere to, they are:

1. Emphasize the largest salient objects
2. Uniformly highlight whole salient regions
3. Establish well-defined boundaries of salient objects
4. Disregard high frequencies arising from texture, noise, and blocking artifacts
5. Efficiently output full resolution saliency maps

To meet all of the aforementioned requirements, the authors define their saliency map function as:

$$S(x, y) = |I_\mu - I'(x, y)| \tag{2.11}$$

where I_μ is the mean image value and $I'(x, y)$ is the corresponding image pixel value in the Gaussian blurred version of the original image. Blurring the image using a Gaussian kernel allows high frequencies to be disregarded (requirement 4) and hence produces a salient regions map. For an in-depth discussion on the process of Gaussian blurring, see Appendix B. Since no downsampling occurs, the yielded saliency map is of the same resolution as the original image (requirement 5).

To include colour information, it is suggested by the authors that Equation 2.11 be implemented in the *Lab* colour space where each pixel location would be a $[L, a, b]^T$ vector and the L_2 norm is the Euclidean distance.

The Frequency Tuned Method Using Wavelets

Ngau et al. later [30] proposed that instead of blurring the image to remove high-frequency noise and artifacts, one could perform wavelet decomposition to yield a down-sampled approximation of the image. This would allow the image to be reconstructed without really

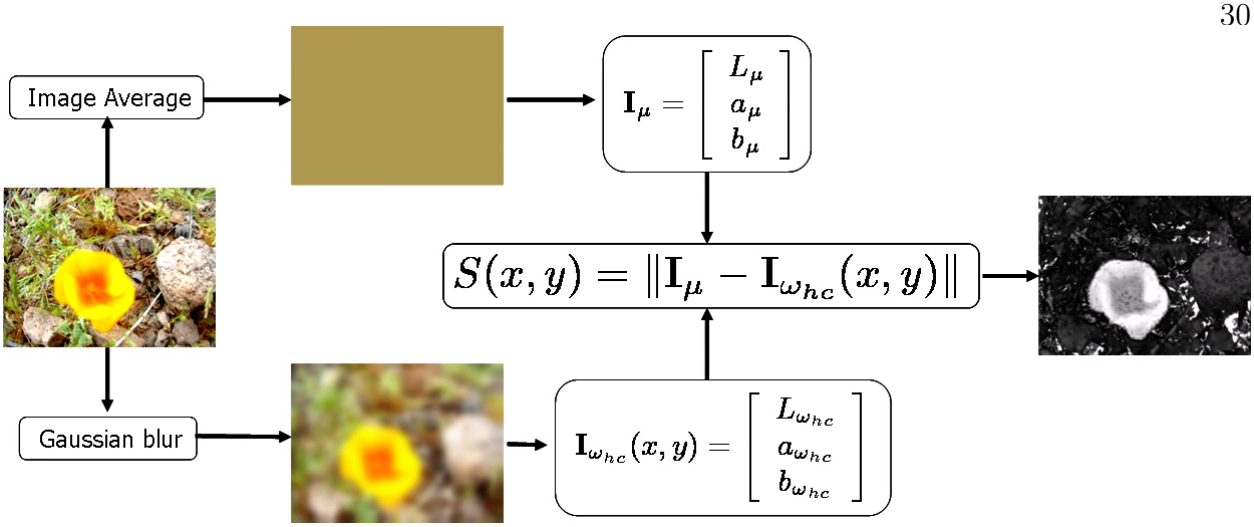


Figure 2.13: Block diagram of frequency tuned method [6].

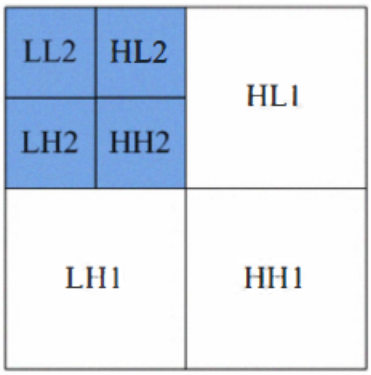


Figure 2.14: Two-level wavelet decomposition.

losing any information, and without resorting to blurring the image which, in a sense, is a de-resolution operation and results in loss of salient objects that are very small in size.

The authors applied the same formula as Achanta et al. (Equation 2.11) except that they applied it to the approximation of the wavelet decomposition (using only one level). They then reconstructed the image back to its original size. To perform wavelet decomposition, a Le Gall 5/3 wavelet is used.

As shown in Figure 2.14, the evaluation of Equation 2.11 is applied to the entire blue area, which is the LL1 sub-band. LL1 refers to the approximation of the first-level decomposition.

To incorporate colour information, it is proposed that this process be applied to each

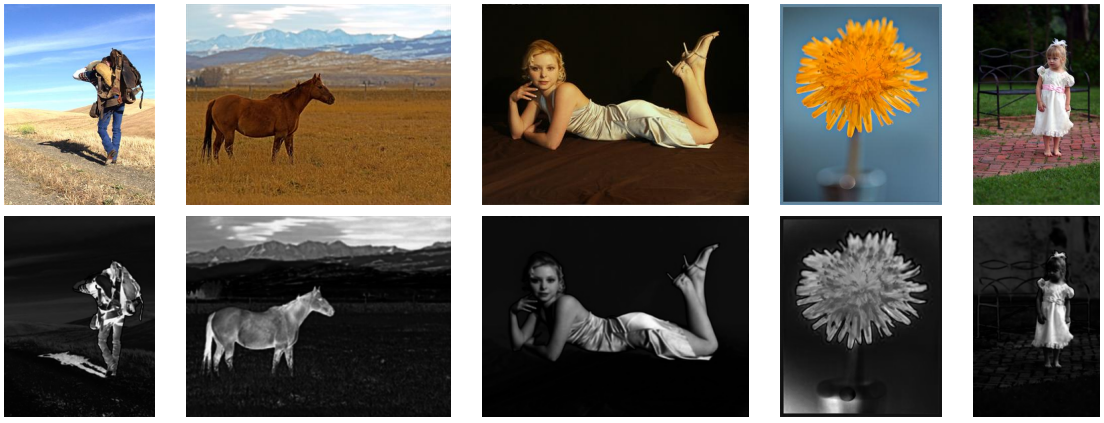


Table 2.4: Samples of the frequency tuned method.

channel of the YCbCr colour space.

2.3.1 Sample Results

Some results are shown to give a better understanding of the frequency tuned method.

It should be noted that in comparison the Itti's method and the SR method, the importance of preserving resolution is very significant. Although Achanta's method does a good job at locating and extracting the region of attention, certain background objects and artifacts are easily picked up. This is shown in the two leftmost images (the walking man and the horse).

2.4 Summary

This chapter discusses the literature with regards to both salient region detection as well as salient edge detection.

Past methods are discussed chronologically. First, the HVS-based method discovered by Itti et. al is discussed in detail. Beginning with the idea of feature integration and then feature selection, the centre-surround operator is discussed. An explanation is given as to how the centre-surround operator traverses a certain feature to extract salient areas in an

image. After this, Itti's normalization operator is discussed as this is also a vital component to this method.

Next, the Spectral Residual method is discussed in detail. The importance of computationally-driven methods is discussed in this section also since this method is the first to focus on computational efficiency. An important concept is discussed in this section, which is that of reconstructing a phase/frequency spectrum without any magnitude spectrum to yield a saliency map (adjustments can be made to extract either regions or edges).

The final method discussed in the salient region detection literature is the Frequency Tuned method. Particular stress is given to the requirements presented by Achanta et al in this work as it sets a new standard for saliency maps. The relationship between this method and the SR is also discussed.

Finally, some light is shed on a modification that can be made to the SR method to allow it to extract salient edges rather than salient regions. This modification is simply to perform the algorithm without the pre-processing step of downsizing the image to a size of 64×64 pixels.

Chapter 3

Novel Advancements

THE novel advancements in this thesis revolve around the discrete wavelet transform (DWT). This chapter begins with an in-depth exploration of the DWT in order to give the reader a thorough understanding of the algorithms that are proposed. The explanation begins with simple one-dimensional signal examples and then continues into more advanced, more applicable two-dimensional image examples.

The main component of this chapter is the explanation of the proposed algorithms, namely the high definition human visual system method, the wavelet residual method, and the frequency-tuned wavelet residual method.

3.1 The Discrete Wavelet Transform

A walk through of the development of DWT is given in this section, starting off with the Fourier transform (FT) which is assumed to be well known to the reader. For simplicity's sake, the discussion will revolve around a 1D signal, and the transition to 2D (i.e. an image) will appear when necessary.

Using the FT, it is well established that any signal can be decomposed into a series of complex exponentials (sines and cosines). However, this limits the decomposition to pure frequency information, so it can be pointed out which frequencies are present in the signal, but not exactly where these frequencies are located in the signal.

A solution to this drawback is the short-time Fourier transform (STFT). A small window

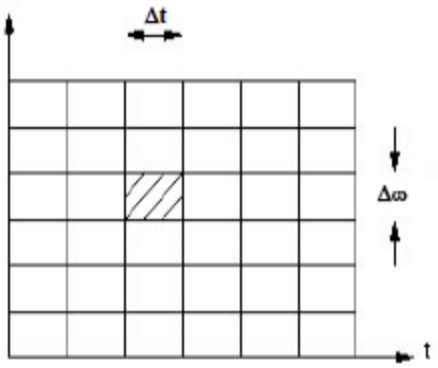


Figure 3.1: Short-time Fourier transform.

is applied to the signal and the FT is applied within that small window, allowing a localization of the frequencies in that area. The window is then shifted throughout the entire signal so that frequencies can be located at their spatial location. Mathematically, the STFT is defined by Equation 3.1:

$$STFT(t, f) = \int x(t + \tau)w(\tau)e^{-j\omega t}dt \tag{3.1}$$

where $x(t)$ is the signal and $w(t)$ is the window that is traversing the signal. The STFT can be illustrated as a tiling of the time-frequency plane, as is shown in Figure 3.1.

There is still a major drawback with STFT. Ideally, the window $\Delta\omega$ would be as small as possible, in order to analyze higher frequencies. Simultaneously, it is ideal that Δt would be as small as possible, in order to get finer spatial representations of the frequencies. However, the two ($\Delta\omega$ and Δt) and inversely proportional, so when one increases, the other decreases.

This is where the DWT comes into play. Although the DWT also incurs the same time/space vs frequency trade off, it allows more flexibility by allowing a wide range of frequencies to be observed over the same spatial location. The wavelet decomposition process is essentially a series of recursive high-pass and low-pass filters that are applied, in the time-domain, to a signal. Each time the signal is split into a high-pass frequency component and a low-pass frequency component, the low-pass component is filtered again with both filters. Therefore, initially, the signal is split at a certain frequency (the cut-off frequency for

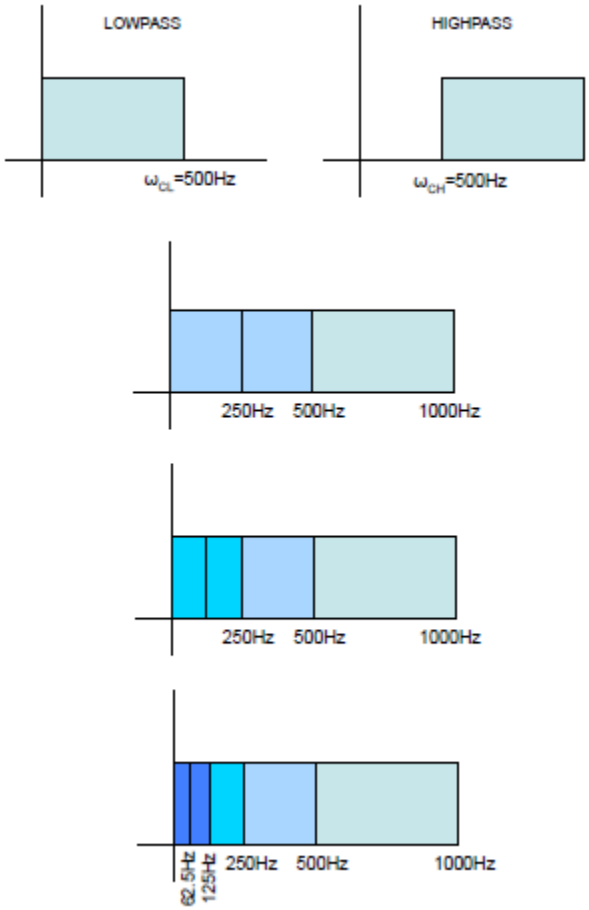


Figure 3.2: Wavelet decomposition process.

both the high-pass and low-pass filters), and then a further division of the low-pass portion into two more frequency bands, then two more, then two more, and so on and so forth. This process is shown in Figure 3.2, where a signal containing frequencies from 0-1000Hz is initially split at 500Hz, then the lower group of frequencies is split at 250Hz, then 125Hz, then 62.5Hz.

With each iteration of this process, the lower group of frequencies is more representative of the signal than the higher band since it contains more information. This will be more evident when considering images.

A numeric example will clarify this process before continuing into the 2D domain of

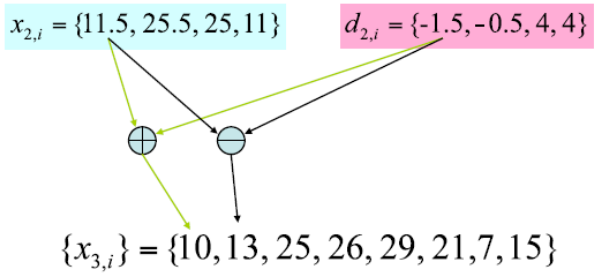


Figure 3.3: Wavelet reconstruction process.

images. Consider the following signal:

$$x_{0,i} = \{10, 13, 25, 26, 29, 21, 7, 15\}$$

If we apply the following transforms (LP and HP respectively):

$$x_{n-1,i} = \frac{x_{n,2i} + x_{n,2i+1}}{2}$$

$$d_{n-1,i} = \frac{x_{n,2i} - x_{n,2i+1}}{2}$$

we get the following components:

$$x_{1,i} = \{11.5, 25.5, 25, 11\} \ , \ d_{1,i} = \{-1.5, -0.5, 4, 4\}$$

where $x_{1,i}$ is the approximation of the signal (since it contains more information), and $d_{1,i}$ is the detail component (since it contains the leftover details of the signal). Breaking the approximation ($x_{1,i}$) down further, we get the following approximation and detail components:

$$x_{2,i} = \{18.5, 18\} \ , \ d_{2,i} = \{-7, 7\}$$

To reconstruct the original signal, we simply combine the approximation and detail components. This is illustrated in Figure 3.3.

Taking this process to the 2D domain, the signal is first decomposed horizontally, and then vertically. This results in a horizontal detail component, and vertical detail component, and

Figure 3.4: Wavelet decomposition example: Lena.

a diagonal detail component (the result of both horizontal and vertical filtrations combined). This is illustrated in Figure 3.4, which shows two levels of decomposition.

Wavelet decomposition locates frequency information at its spatial location. Because of it's lossless resizing power, wavelets are used in JPEG2000 compression.

One of the most commonly used wavelets is the Haar wavelet (Figure 3.5), which was used for all the algorithms in this thesis. This kernel traverses the image horizontally to obtain the horizontal detail component of the wavelet. It is then applied to the image vertically to obtain the vertical component, and finally both horizontally and vertically for the diagonal component. The remaining image is the approximation component.

3.2 Application to Region Detection

3.2.1 The High Definition Human Visual System Model

There are two main drawbacks with Ittis method in that it requires a lot of processing time (so it cannot be applied in real time), and it is of very low resolution (so an exact outline of the object cannot be extracted). The goal of this method is to overcome these drawbacks

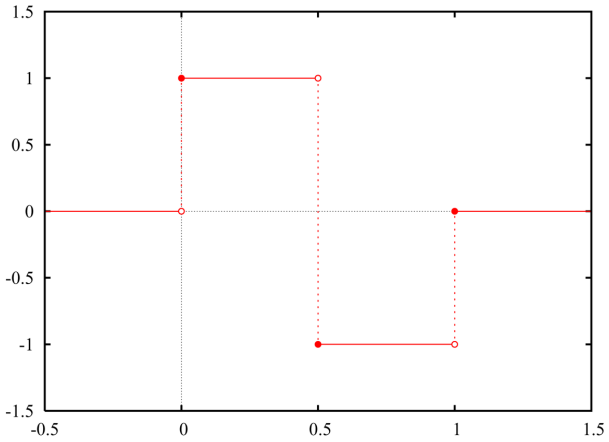


Figure 3.5: The Haar wavelet.

while still including the feature integration theory to allow for a biologically inspired process in detecting salient regions.

Wavelet theory has recently taken on a large role in image processing due to its lossless resizing capabilities. The one factor eluding to loss of resolution in Ittis method is the downsampling using Gaussian pyramids. As such, the wavelet decomposition is used in this work, specifically, a one-level-down approximation is taken as the downsample to the original image. Performing the centre-surround function using the original image and said approximation, we have the capability to resize the resulting saliency map back to the size of the original image without any loss in resolution.

Figure 3.6 shows the block diagram of the proposed method. It should be noted that Figure 3.6 only shows the grayscale component as well as the red-green color opponency, and not the blue-yellow opponency. This was done to save space. The blue-yellow opponency block diagram is identical to the red-green one except it uses the blue and yellow color maps rather than the red and green maps.

The centre-surround used in the proposed method is the same used by Itti. Taking the wavelet approximation as the centre scale, C(θ), and the original map as the surround scale, S(θ), the centre-surround function is implemented by downsizing the surround scale to the size of the centre-scale using bicubic interpolation, and performing absolute point-

by-point subtraction. It is now clear why wavelets are necessary, since the output of the centre-surround function is not the same size as the original image.

In Figure 3.6, the input image is first put through a set of linear filters to extract grayscale information (Equation 2.2) as well as red, blue, green, and yellow information (Equations 2.3, 2.4,2.5, 2.6). Each of these information maps are then decomposed using a wavelet filter to extract an approximation as well as horizontal, vertical, and diagonal detail maps. With regards to the grayscale map, the wavelet decomposition components are labeled GrA (approximation), GrH (horizontal), GrV (vertical), and GrD (diagonal). The same applies to the red and green maps, except the prefix Gr (gray) is replaced with R (red) and G (green), respectively. The centre-surround operation (shown as the θ operator) is performed using the approximation and original information map. Such is the case for the grayscale component. However, to implement color opponency, the centre-surround operation is applied to the red information map with the green wavelet approximation, and vice versa, and similarly with the blue-yellow opponency (again, not shown). The output of the centre-surround operation is then used to reconstruct the image using the wavelet decomposition components. All of the resulting conspicuity maps are then normalized (using Ittis normalization function [3]) and summed to form the final saliency map.

In this method, orientation saliency was not implemented as in Itti's method [3]. This is because at such a high scale (the original and one scale down), an orientation map looks similar to an edge map. This obstructs the final saliency map as the edges in the orientation map may not lay exactly at the boundaries of the regions yielded by the intensity and color maps and hence some objects appear to be surrounded by a ringing/halo effect. Also, the orientation results did not provide much new information anyway, and ignoring them also improves processing time.

3.2.2 Sample Results

Some results are shown to give a better understanding of the HDHVS method. It can be seen that the HDHVS method does a good job at detecting salient objects at many levels.

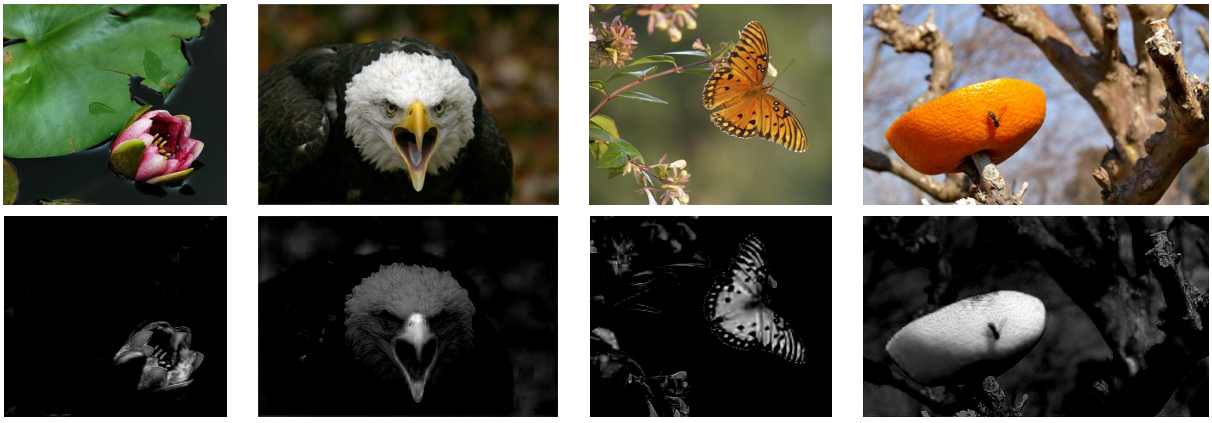


Table 3.1: Samples of the HDHVS method.

The image of the eagle shows that the beak is the most salient object, but that the entire white head is also very salient with respect to the remainder of the image. The importance of colour opponency is also evident in these examples as all the images are full of colour, yet the salient object is still correctly extracted.

In the approaches discussed in Chapter 2 that do not consider colour opponency, white or luminous regions take precedence in the saliency map (e.g. sky or background). Here this is not the case, and the new operator attempts to focus on the more conspicuous region (see rightmost image in Table 3.1).

3.3 Application to Edge Detection

3.3.1 The Wavelet Residual Model

There are two fundamental advantages that the discrete wavelet transform (DWT) has over the Fourier transform. They are: the ability to spatially locate frequency components, and the ability to sub-band the frequency spectrum [6]. The second mentioned advantage allows us to locate salient regions at various levels of saliency. The highest sub-band (largest frequency) yields the most salient regions since the pop-out is largest in magnitude.

The algorithm for the proposed method is as follows:

$$[A(f), H(f), V(f), D(f)] = \mathfrak{W}[I(x)] \tag{3.2}$$

$$S(x) = \mathfrak{W}^{-1}[0, H(f), V(f), D(f)] \tag{3.3}$$

In Equation 3.8, $I(x)$ is the input image (assumed to be already at the desired scale of decomposition), $\mathfrak{W}$ is the wavelet decomposition operator, and $A(f)$, $H(f)$, $V(f)$, $D(f)$ are the approximation, horizontal component, vertical component, and diagonal components of the wavelet decomposition, respectively. In Equation 3.3, $H(f)$, $V(f)$, $D(f)$ are the horizontal, vertical, and diagonal components respectively, and $S(x)$ is the saliency map.

Figure 3.8 shows the saliency maps computed using the first three sub-bands of a sample image. It should be noted that the first sub-band contains the sharpest edges. This is because the frequencies extracted are very high and hence produce a thin edge. The second and third subbands contain salient edges in essentially the same locations as the first sub-bands, but the edges are thicker; however they also contain traces of less-salient edges which may not necessarily be present in the first sub-band.

In contrast to the SR method, this method produces results that contain more information. More times than not, this information is pertaining to a salient edge, but of course, this is not always the case.

3.3.2 Sample Results

Some results are shown to give a better understanding of the wavelet residual method. The main thing to notice in these results is that this operator is not an ordinary edge detector. This is evident because there are edges in the images that are suppressed in the saliency map. A negative aspect to this method is that it tends to over detect edges and falsely classify certain edges as salient. This is definitely something that must be tuned in the future.

Table 3.2: Samples of the wavelet residual method.

3.3.3 The Frequency-Tuned Wavelet Residual Model

Using Achantas requirements as a basis for this method, the requirements that are necessary for this method are preserved, while the others are neglected. From their requirements, as specified in [6], the following three are taken:

1. Establish well-defined boundaries of salient objects
2. Disregard high frequencies arising from texture, noise, and blocking artifacts
3. Efficiently output full resolution saliency maps

This method uses the logic of Achanta(frequency-tuning), Ngau(using wavelets to attain lossless information processing), as well as Ma(frequency-to-time domain reconstruction using only the frequency component).

From Achanta, it is realized that frequency-tuning allows salient regions to "pop out" more (Equations 3.5, 3.6, and 3.6). In order to detect salient edges more clearly, we perform this operation on the frequency components of the wavelet decomposition of the image. From Ngau, it is realized that performing such an operation using wavelets allows for a reconstruction without any loss of information (Equations 3.4 and 3.8). From Ma, it is

realized that reconstructing using only the frequency components yields a map of the salient edges (Equation 3.8). The formulation of this method is as follows:

$$[A(f), H(f), V(f), D(f)] = \mathfrak{W}[I(x)] \tag{3.4}$$

$$H^*(f) = |H_\mu(f) - H'(f)| \tag{3.5}$$

$$V^*(f) = |V_\mu(f) - V'(f)| \tag{3.6}$$

$$D^*(f) = |D_\mu(f) - D'(f)| \tag{3.7}$$

$$S(x) = \mathfrak{W}^{-1}[0, H^*(f), V^*(f), D^*(f)] \tag{3.8}$$

where $I(x)$ is the input image, $\mathfrak{W}$ and $\mathfrak{W}^{-1}$ are the wavelet decomposition and inverse wavelet decomposition operators, and $A(f)$, $H(f)$, $V(f)$, and $D(f)$ are the approximation, horizontal, vertical, and diagonal components of the wavelet decomposition, respectively. $H'(f)$, $V'(f)$, and $D'(f)$ are the Gaussian blurred versions of the horizontal, vertical, and diagonal components, respectively. $H_\mu(f)$, $V_\mu(f)$, $D_\mu(f)$ are the mean pixel values of the horizontal, vertical, and diagonal components, respectively. $H^*(f)$, $V^*(f)$, and $D^*(f)$ are the horizontal, vertical, and diagonal frequency-tuned components respectively, and $S(x)$ is the saliency map.

Applying the frequency-tuning operation while in the decomposed state allows the salient edges to be further accentuated, while suppressing the inattentive points. This is because while high frequency artifacts are removed by the wavelet operator, the most common of the remaining frequencies are made to stand out by means of the frequency tuning operator.

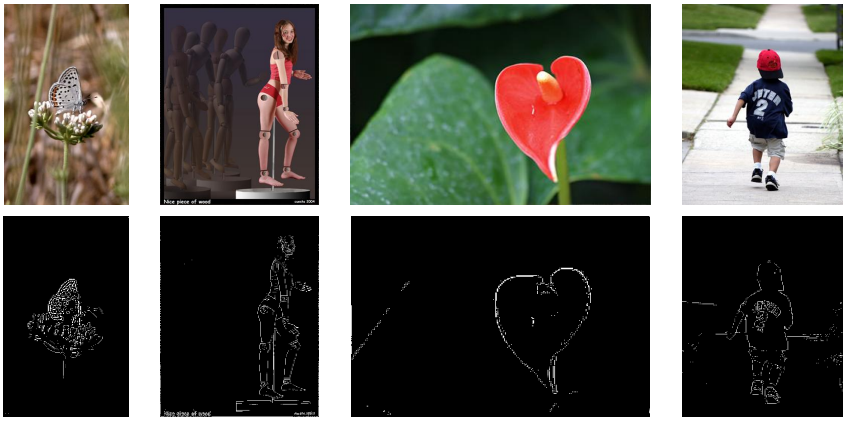


Table 3.3: Samples of the frequency tuned wavelet residual method.

Sample Results

Some results are shown to give a better understanding of the frequency tuned wavelet residual method.

Similar to the results shown in the discussion of the WR, these examples show the FTWR's power to suppress backgrounds and artifacts, although not entirely.

3.4 Summary

This chapter provided an explanation of the novel advancements proposed in this thesis with regard to visual saliency mapping. Since all of the proposed methods depend on the discrete wavelet transform, the chapter begins with an explanation of it.

The first proposed method that is described is the high resolution human visual system based method. This method is a direct offspring of Itti's human visual system based method in that it employs the same concepts of centre-surround and feature integration. However, in order to gain the benefits of high resolution, wavelets are introduced into the algorithm as a replacement to Gaussian pyramids. The lossless resizing power of wavelets allow scales of the image to be created without any loss of information upon resizing back to the original size. A block diagram is shown in Figure 3.6.

The second proposed method that is described is the wavelet residual method. This method is a direct offspring of the spectral residual method in that it employs the same concept of decomposing an image and then reconstructing it using only the frequency information. In [5], by contrast, the image was decomposed using the Fourier transform, and when it is recomposed using only the frequency information, the resulting saliency map was forced to be very impulsive (contains many unconnected dots). Performing the same process but decomposing using wavelets rather than FT allows for a smoother reconstruction, as is evident in the sample results.

The final proposed method that is described is the frequency tuned wavelet residual method. This method is a direct offspring of the frequency tuned method as well as the spectral residual method in that it employs the concepts of decomposing and reconstructing using only the frequency information, as well the frequency tuning operation. The image is once again decomposed using wavelets, but before reconstruction, the detail components (vertical, horizontal, diagonal) of the wavelet are frequency tuned as per Equation 2.11.

Sample results are shown for each process after the explanation is given. For more results, as well as comparisons, see Chapter 4.

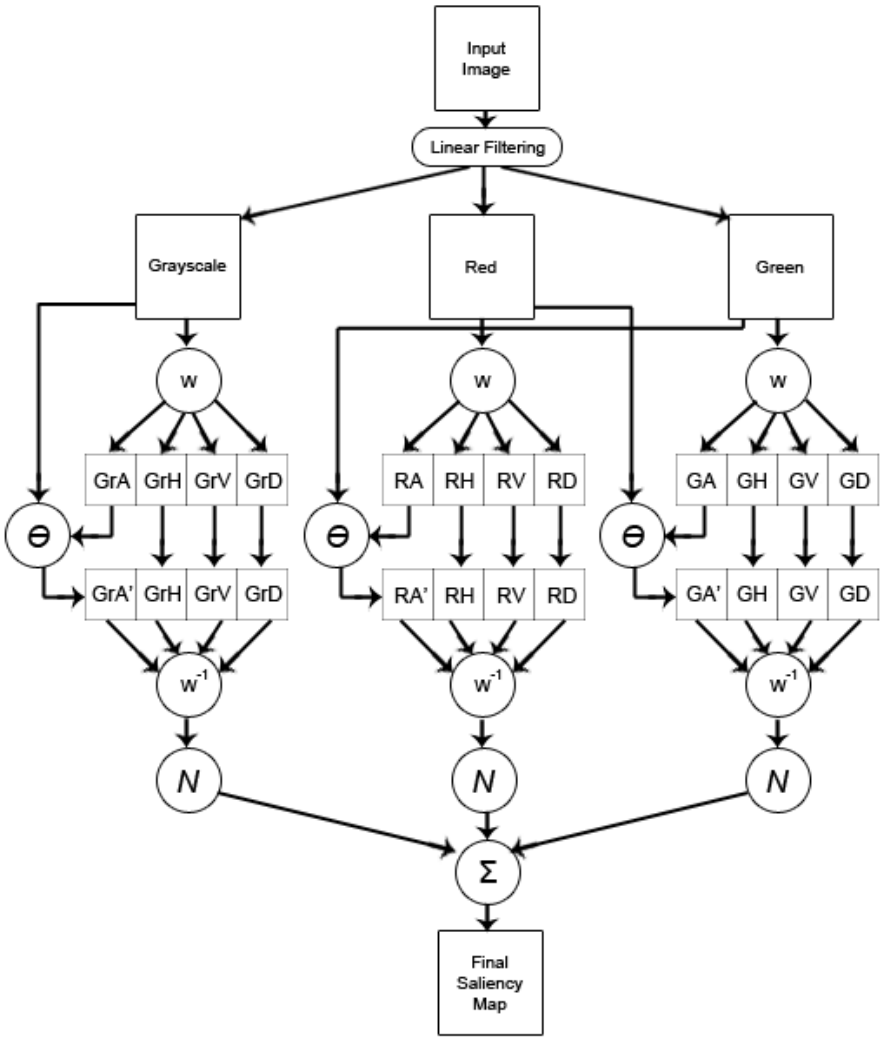


Figure 3.6: Block diagram of the proposed HDHVS model.

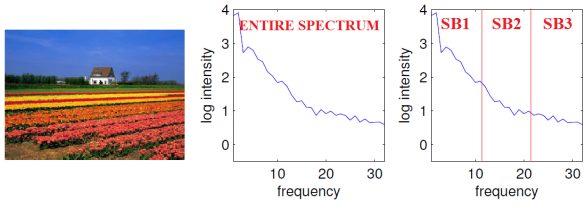


Figure 3.7: Spectrum breakdown.

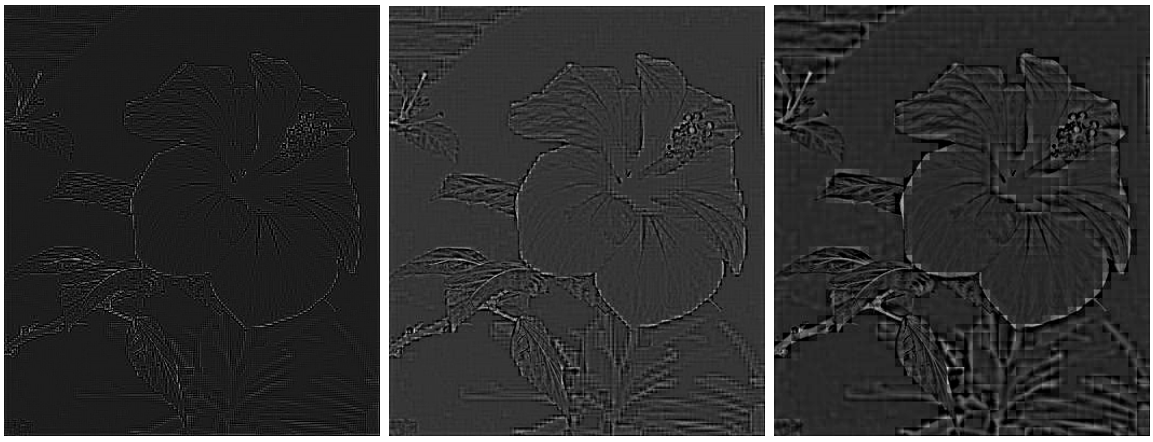


Figure 3.8: Saliency maps using first three sub-bands.

Chapter 4

Results

THE saliency detection methods presented in this thesis have been tested on a variety of images containing a wide and varying array of features such as intensity, colour, general orientation, and texture, to name a few. For a thorough analysis of the results, sets of images were tested together for a fair comparison. The test images are categorized as images of flowers, images of cars, images of human figures, images of signs, images of animals, and miscellaneous images. These specific categories were chosen because each has a unique set of features that can show the strengths and weaknesses of each method used.

The tests were run on over 5000 images, all taken from the MSRA image database [31]. The same images were used to test all the methods, for a fair comparison. The images all have dimensions of 300×400, or are very close to that size. All of the images are RGB images, and none are purely grayscale.

The tests are split into region detection and edge detection methods. In the region detection section, our proposed method, HDHVS, is compared to the frequency-tuned method proposed by Achanta, the HVS method proposed by Itti, and the spectral-residual method proposed by Zhang. The edge detection section contains results of the modified spectral-residual method alongside both of our proposed methods, the WR, and the FTWR.

The MSRA database provides a ground truth as an attempt at an evaluation process. However, this ground truth consists of a rectangle identifying the attentive object. This makes comparing results to said ground truth somewhat meaningless because certain situa-

Figure 4.1: Sample ground truth comparisons. From left to right: original image, HDHVS method, Achanta's method, Itti's method, and the SR method.

tions would yield the same outcome, although they are actually very different. To illustrate this point, Figure 4.1 shows an image and a ground truth comparison of the results of the methods in this thesis. It can be seen that each method has the majority of its pixels within the boundaries of the red rectangle (indicating the location of the attentive object). This comparison is not plausible because clearly having the object within the limits of some boundaries does not conclude any measures of precision, resolution, or strength of saliency.

The tests were run using Matlab R2010b on a Dell XPS M1330 laptop with a T5750 2GB Intel Core 2 Duo processor, 3GB of RAM, and Windows 7.

4.1 Salient Region Detection Results and Comparisons

As a preliminary observation, the reader should notice that the results of the HDHVS method are most comparable to those of Achanta's method, since they both follow the same set of regulations (same resolution as original image, remove high frequency noise, etc). Although Itti's method and the SR are not very comparable, they are still presented for reference.

For these tests, original authors' implementations were used when possible. For Achanta's method, his implementation was available on his website, and hence that is the implementation that was used. For Itti's method, the implementation on his website is in the C++ language, so Dirk Walther's Matlab implementation was used, which is available at www.saliencytoolbox.net. The spectral residual method was not available from the original authors so it was implemented by the author of this thesis; this implementation was used for these tests.

All of the implementations are programmed as they are specified in this thesis. For a fair

comparison, all of the results are normalized to a dynamic range of [0,1] and are resized to the dimensions of the original image.

When looking at the results, the reader should keep in mind that a saliency map is a grayscale image containing pixels ranging from a value of 0 to a value of 1. The darker (closer to absolute black) a pixel is, the less salient it is; and vice versa; the lighter (closer to absolute zero) a pixel is, the more salient it is.

Figure 4.2 shows the results of the salient region detectors applied to various images of flowers. Figure 4.2(b) shows that the flower petals have been correctly detected by the HDHVS method, but that an artifact has also been detected. This artifact is another plant, but it blends with the rest of the greenery and does not stand out as much as the yellow petals. Figure 4.2(c) shows that Achanta's method detected the yellow petals, but also all of the green stems. It also does not show a distinction in saliency between the stems and the petals as the salient regions detected are all roughly of the same strength. Figure 4.2(d) shows that Itti's method detected the location of the center of the flower, but other artifacts were also falsely detected to be salient. The SR method shows random points being detected. Figure 4.2(e) shows points on the left side of the image, which is where the flower is, but no shape is really formed. The general trend in the rest of Figure 4.2 is that the HDHVS tends to detect the flower petals very well, and also tends to suppress all other aspects of the image. This is perhaps because of the strong presence that petals have in the image. Achanta's method also detects petals well, but not as strongly as the HDHVS, and also detects artifacts. This is an issue only because this method tends to give the artifacts the same level of saliency as the petals. Itti's method and the SR both locate the flower, but other artifacts are almost always present (see Figures 4.2(i) and 4.2(y)). A particular point to note is that Figure 4.2(v) shows that the HDHVS is the only method to detect the entire flower in Figure 4.2(u). The rest of the methods tend to focus on the white stripe.

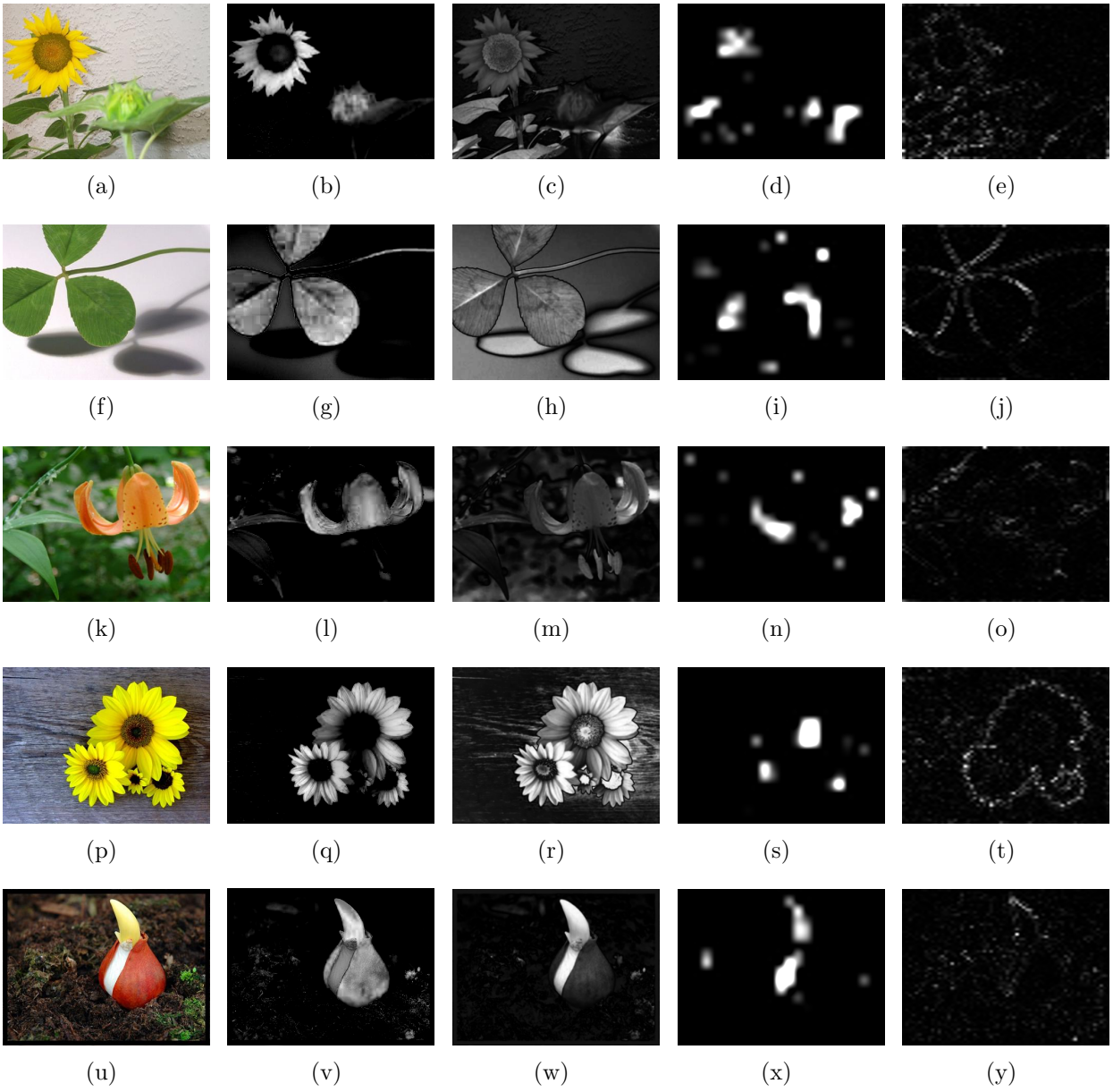
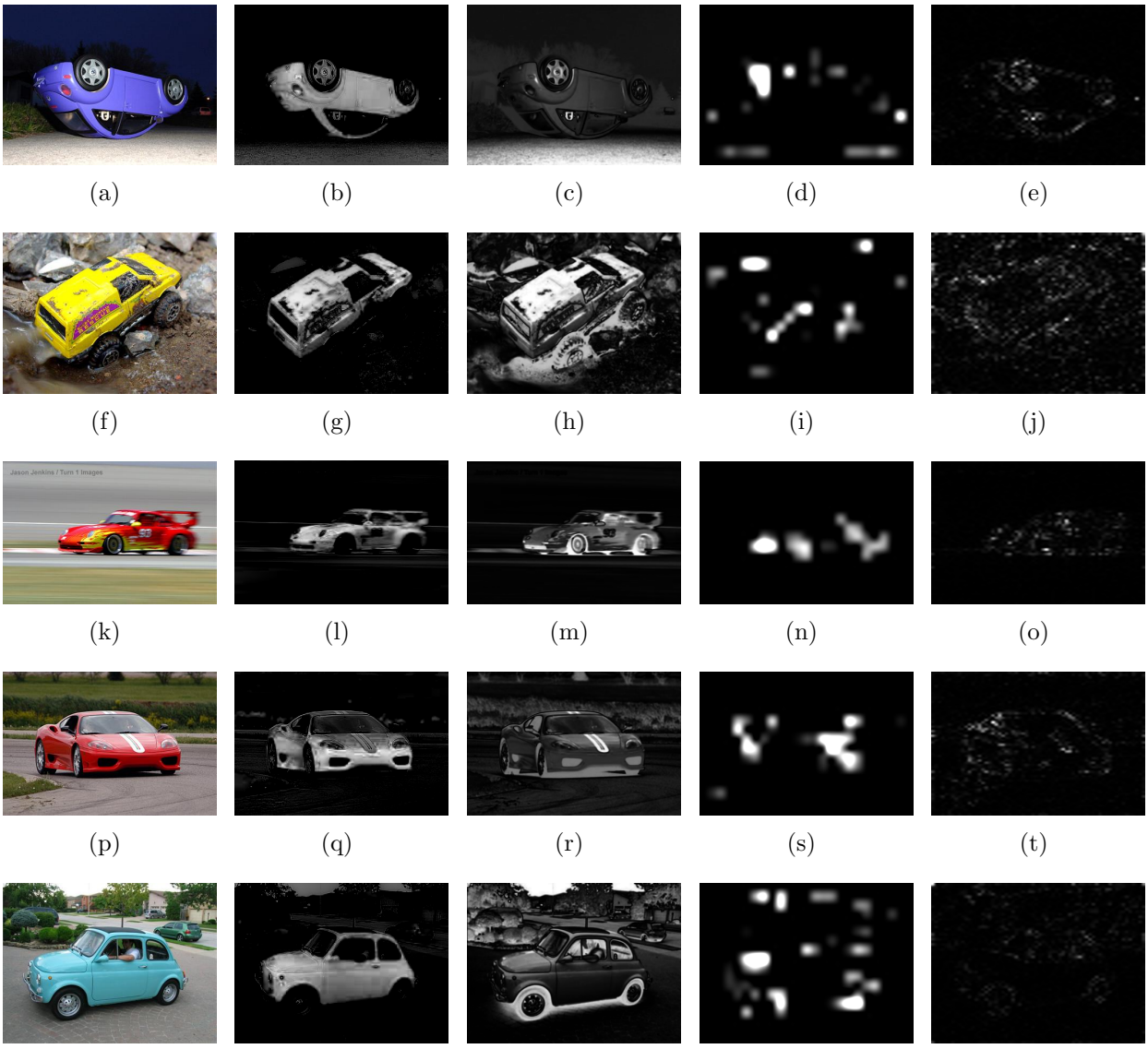


Figure 4.2: Comparison of flower image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

Figure 4.3 shows that the HDHVS tends to detect the car's body over anything else. A drawback of this is that other components of the car (such as the windshield, windows, head/tail lights and tires) are not detected to be salient. Achanta's method usually picks up the body, but not as strongly as the HDHVS. However, Achanta's method does a good job

at picking up the other components of the car that the HDHVS tends to miss (windshield, tires, etc). A negative aspect to this method is that it also falsely classifies parts of the background as salient (see Figures 4.3(c), 4.3(h), and 4.3(w)). Itti's method and the SR method, as in Figure 4.2, detect the location of the car, but Itti's method yields a blurry spot while the SR method yields a blurry outline.



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

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

(p) (q) (r) (s) (t)

(u) (v) (w) (x) (y)

Figure 4.3: Comparison of car image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

Figure 4.4 shows that the HDHVS tends to detect human skin as salient, and also tends to not detect clothing in the presence of a lot of human skin. There are cases (such as Figures 4.4(g) and 4.4(q)) where both the skin and the clothes are detected. Figures 4.4(b), 4.4(l), and 4.4(v) are good examples of the ability of the HDHVS to detect human skin. In contrast to this, Achanta's method tends to pick up clothing more than human skin, such as in Figures 4.4(r) and 4.4(w). The contrast is very vivid between Figures 4.4(v) and 4.4(w) in that the HDHVS strongly detects the kin and neglects the clothes, and Achanta's method strongly detected the clothes and neglects the skin. Itti's method detects the locations of the human figures although Figure 4.4(s) does a very poor job of even that. The SR method does a rather good job at locating the human figure and suppressing the background. This is particularly true in Figure 4.4(y).

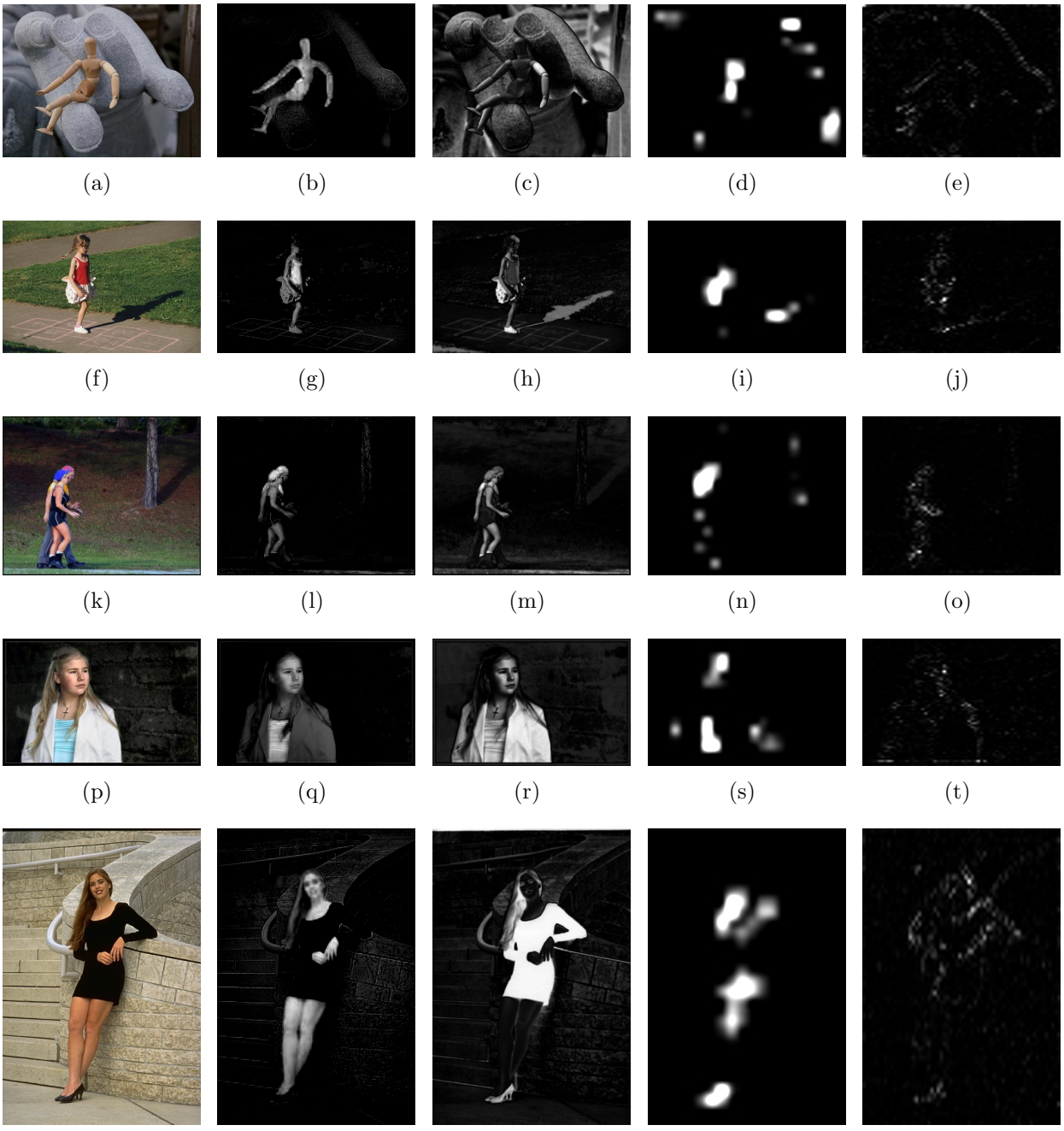
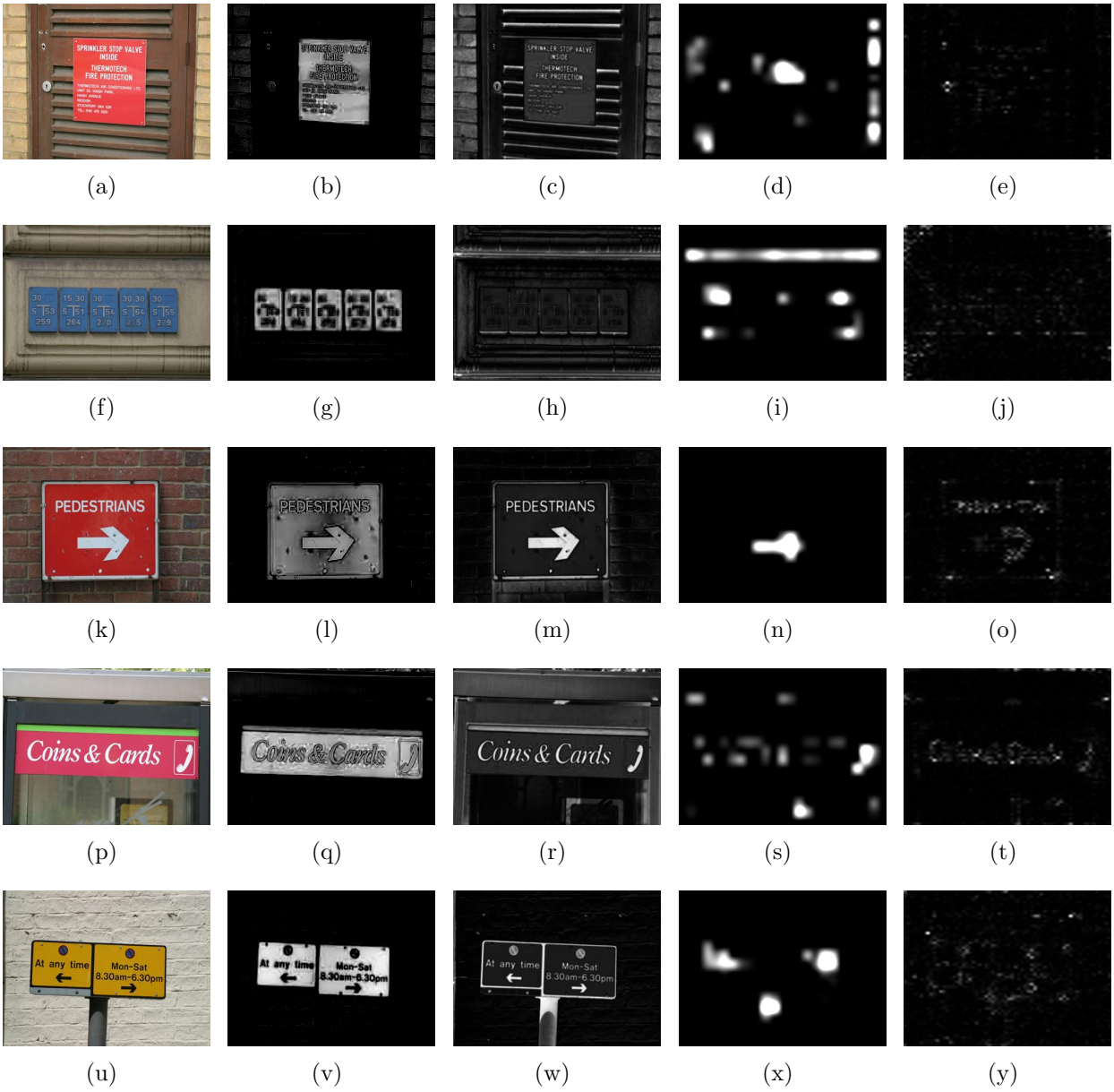


Figure 4.4: Comparison of human image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

Figure 4.5 shows that there is a drastic difference between the HDHVS and Achanta's method. The HDHVS detects the coloured region of the sign (the backdrop) as the most

salient region, which Achanta's method detects the lettering to be the most salient region. In Figures 4.5(l) and 4.5(q), the HDHVS picks up both the backdrop as well as the lettering of the sign; however, there is a clear outlines of the letters. Achanta's method picks up the letters on the sign regardless of their colour (compare Figure 4.5(m) to Figure 4.5(w)). Itti's method picks up the shapes on the signs when they are big enough (see Figures 4.5(n) and 4.5(s)). The SR method has a very difficult time distinguishing the signs from the background. The saliency maps yielded by the SR method look like noise.



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

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

(p) (q) (r) (s) (t)

(u) (v) (w) (x) (y)

Figure 4.5: Comparison of sign image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

Figure 4.6 shows a very important aspect of the HDHVS method: it's ability to detect various levels of saliency. In Figure 4.6(w), Achanta's method detects the teddy bear'd body as the most salient object in the image, and completely overlooks everything else. The HDHVS on the other hand, detects the rose as the most salient object, and the teddy bear

as the next-most salient object, displaying it in a lighter shade of grey. Another important feature shown in Figure 4.6 is the apparent difference in the approaches between the HDHVS and Achanta's method. Figures 4.6(l) and 4.6(m) show the respective saliency maps of the HDHVS and Achanta's method of Figure 4.6(k). The HDHVS method detected the beige coloured dog, while Achanta's method detected the black coloured dog.Also, in Figures 4.6(c), 4.6(h), and 4.6(r), Achanta's method shows the darkest parts of the image as the most salient object, when this is clearly not the case. Itti's method locates correct objects in only two out of the five examples shown; Figures 4.6(n) and 4.6(x), but even these are a bit of a stretch. The SR method decently yields the locations of the objects with respect to the same images as Itti's method (Figures 4.6(o) and 4.6(y)).

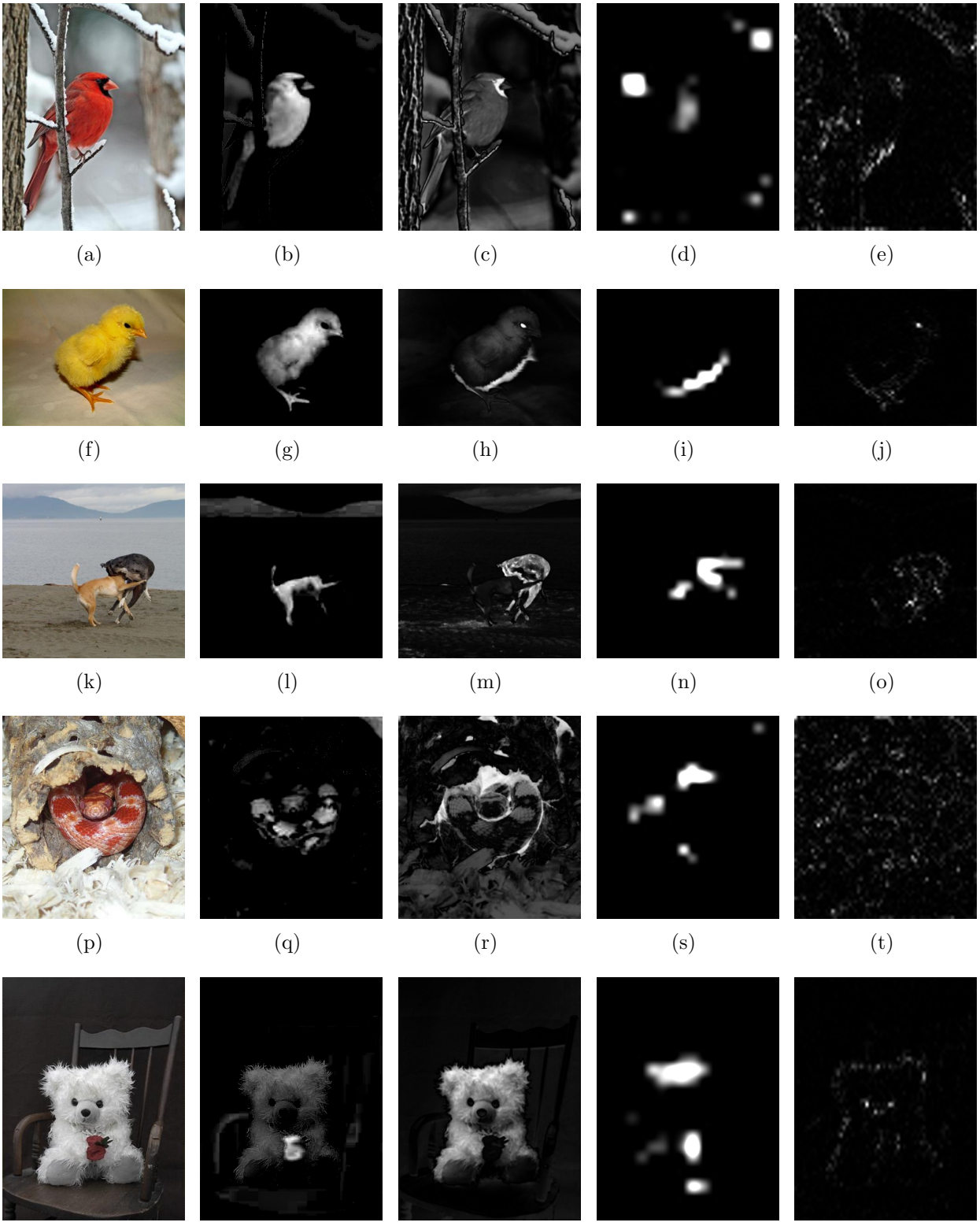


Figure 4.6: Comparison of animal image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

The results in Figure 4.7 show the importance of colour in saliency detection. Although Achanta's method has an aspect of colour to it, it does not explicitly try to find saliency in colour like the biologically inspired methods do (HDHVS and Itti's method). Aside from Figure 4.7(a), all of the images in Figure 4.7 have very strong colour aspects to them. The HDHVS method picks up all of the attentive objects while Achanta's method either neglects them entirely (Figures 4.7(h) and 4.7(m)) or detects them as being equally salient to background objects (Figures 4.7(r) and 4.7(w)). Also, Figure 4.7(b) shows great detail, as well a suppression of the shadows in the image, while Figure 4.7(c) shows that Achanta's method picks up the pencils as well as the shadows. Although Itti's method presents its usual weaknesses, a point to notice is in Figure 4.7(n), where none of the letters are detected, but the location of the toy pig is. The same property is present (or lacking rather) in the SR method, where Figure 4.7(o) does not detect the pig at all, but rather picks up all of the letters. This further proves the importance of colour in saliency detection.

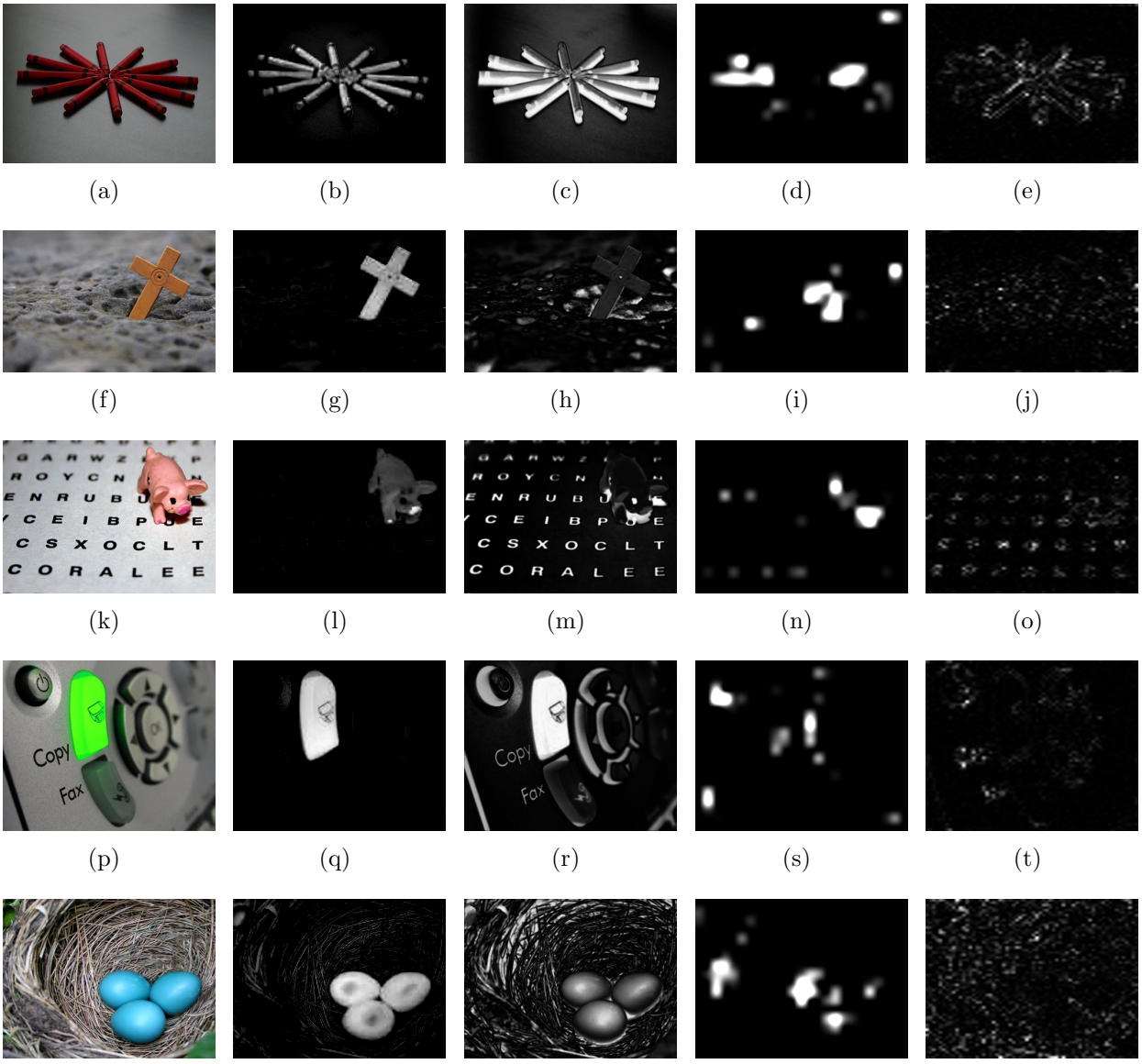


Figure 4.7: Comparison of miscellaneous image results using salient region detectors. From left to right: Original image, HDHVS results, Achanta's results, Itti's results, SR's results.

4.2 Salient Edge Detection Results and Comparisons

Since the concept of edge detection is novel, only the modification to the SR is used for comparison purposes. In general, the WR and FTWR tend to pick up objects more completely than the SR method does. However, the WR and FTWR are very similar. This may lead

to a conclusion that perhaps the FTWR needs some more modifications to perform better, or perhaps some variable to be included that adjusts based on the properties of the image.

Since these saliency maps are meant to emphasize edges, a normalization process of converting pixels values to a range of [0,1] is not realistic. This is due to the fact that there are some pixels already at value 0 and some already at value 1. However, most pixels are somewhere else in this range. To overcome this issue, the results were thresholded using Otsu's method, which locates the optimal threshold value. This process allows for comparable saliency maps.

All of the implementations are programmed as they are specified in this thesis.

When looking at the results, the reader should keep in mind that a saliency map is a grayscale image containing pixels ranging from a value of 0 to a value of 1. The darker (closer to absolute black) a pixel is, the less salient it is; and vice versa; the lighter (closer to absolute zero) a pixel is, the more salient it is.

Figure 4.8 shows a sample set of images and their results using the salient edge detectors mentioned in this thesis. The first two test images (Figures 4.8(a) and 4.8(e)) are examples that show that the WR and FTWR extract more salient content than the SR. In the case of Figure 4.8(a), more petals are detected in the WR and FTWR than in SR, and in the case of Figure 4.8(e), the petals, although not complete, are more well defined. The latter two test images (Figures 4.8(i) and 4.8(m)) are examples that show that the WR and FTWR can sometimes over detect salient edges, to the point where the salient object no longer stands out. This is more true in the case of Figure 4.8(i) than in Figure 4.8(m). Figure 4.8(i) shows the flower being detected but also many of the surrounding leaves and stems, which pollute the saliency map, removing the flower from the spotlight. The saliency maps yielded from Figure 4.8(m) on the other hand, are cluttered from within the petals, picking up what appears to be noise.

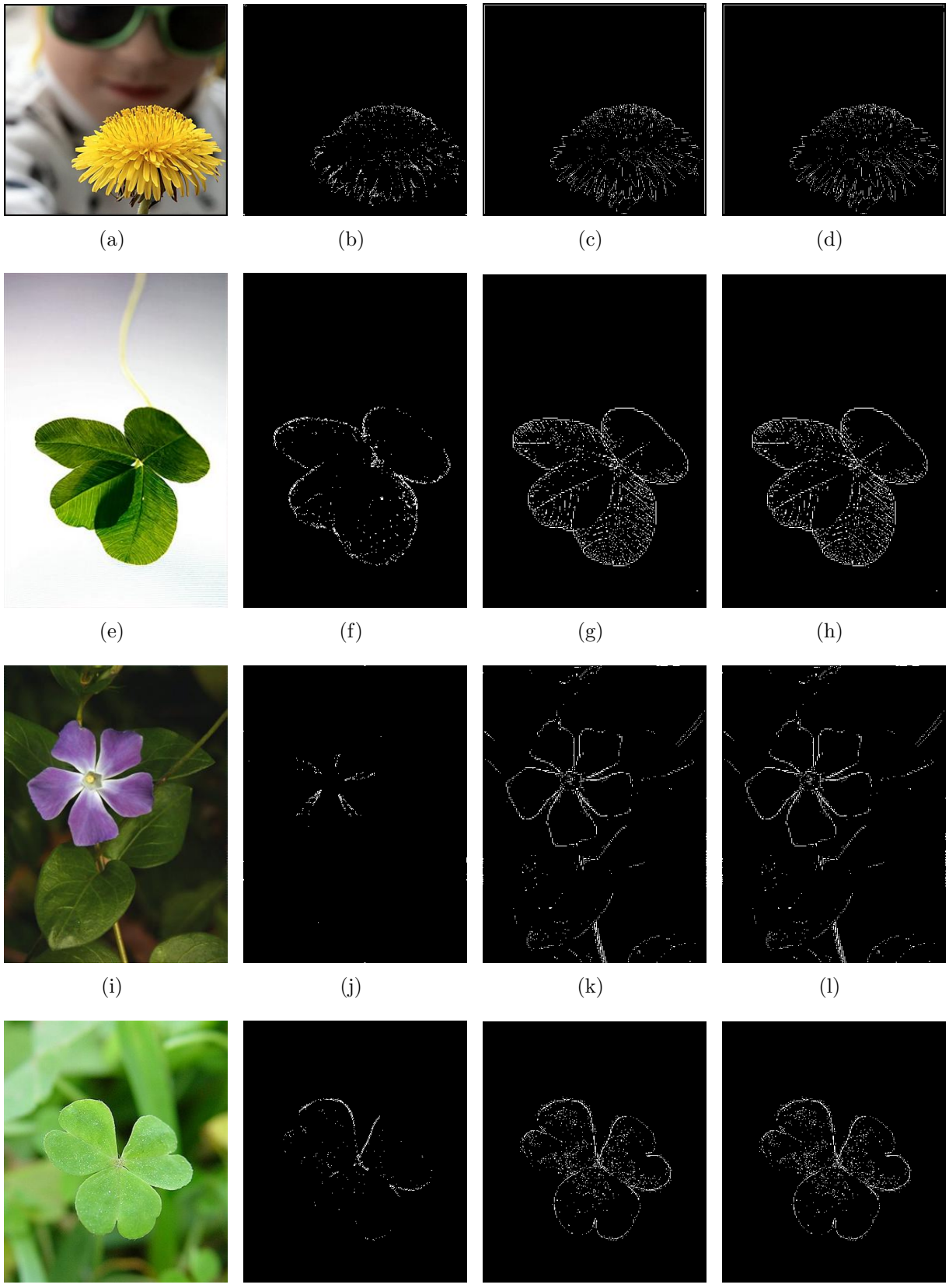


Figure 4.8: Comparison of flower image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

In contrast to the images in Figure 4.8, the results in Figure 4.9 show samples yielding much better results. In all the cases in Figure 4.9, SR returns an incomplete object, while the WR and FTWR return almost perfectly constructed objects, with little to no background artifacts present. In Figures 4.9(k) and 4.9(l), it can be argued that the leaves surrounding the flower should not be present in the saliency map, and this can be understood to be another case of the WR and FTWR over detecting salient edges. Also, in Figures 4.9(o) and 4.9(p), it can be argued that the stem should not be in the saliency map.

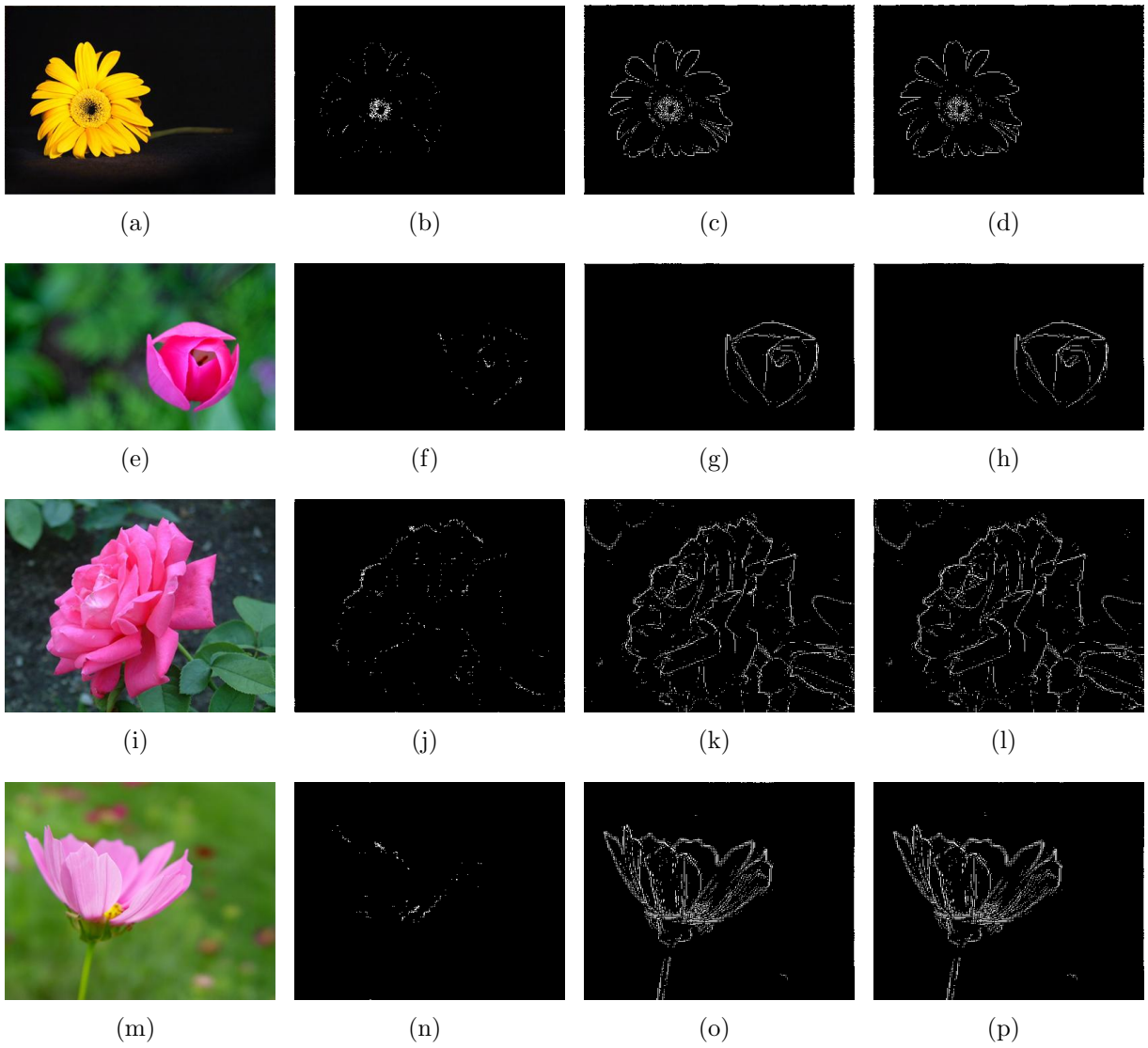


Figure 4.9: Comparison of more flower image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

Figure 4.10 again shows the ability that both the WR and the FTWR possess to over detect salient objects. In all of the results yielded by the SR, only the car is detected, although the car is not entirely outlined. The results yielded by the WR and the FTWR show most of the car, but background noise is also picked up (parts of the road/race track). This shows that the WR and FTWR perhaps require some fine-tuning, or require some parameter to be adjusted based on some property of the image.

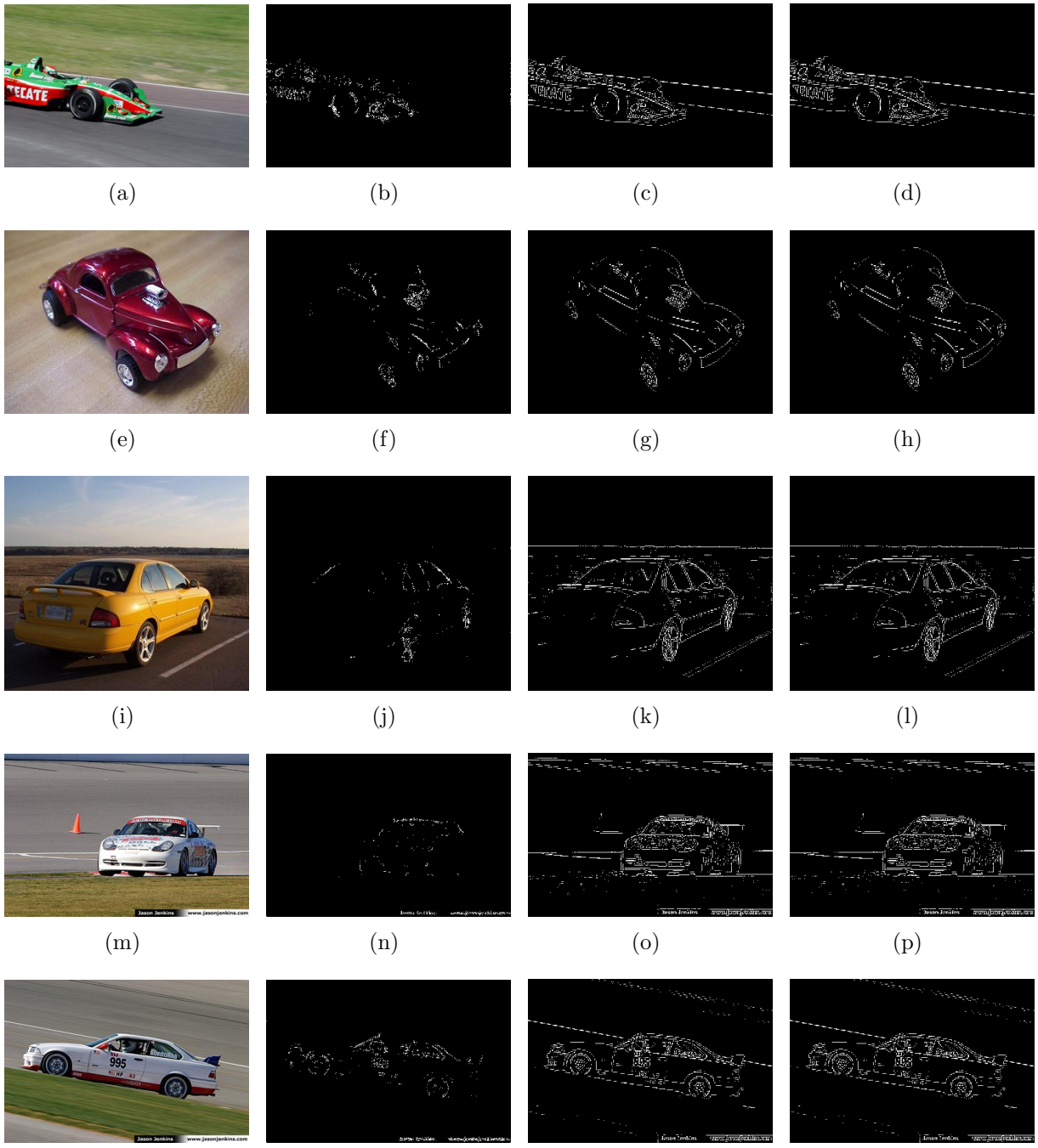


Figure 4.10: Comparison of car image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

Figure 4.11 shows the WR's and FTWR's ability to detect faces in an image. Along with

the SR method, the WR and FTWR pick up bodies well in general, but the WR and FTWR do a much better job at outlining details of the human face. This is evident mostly in the results yielded from Figures 4.11(e), 4.11(i), 4.11(m), and 4.11(q). In these examples, the SR does not pick up any details of the face, while both the WR and FTWR pick up most, or all, of the faces present.

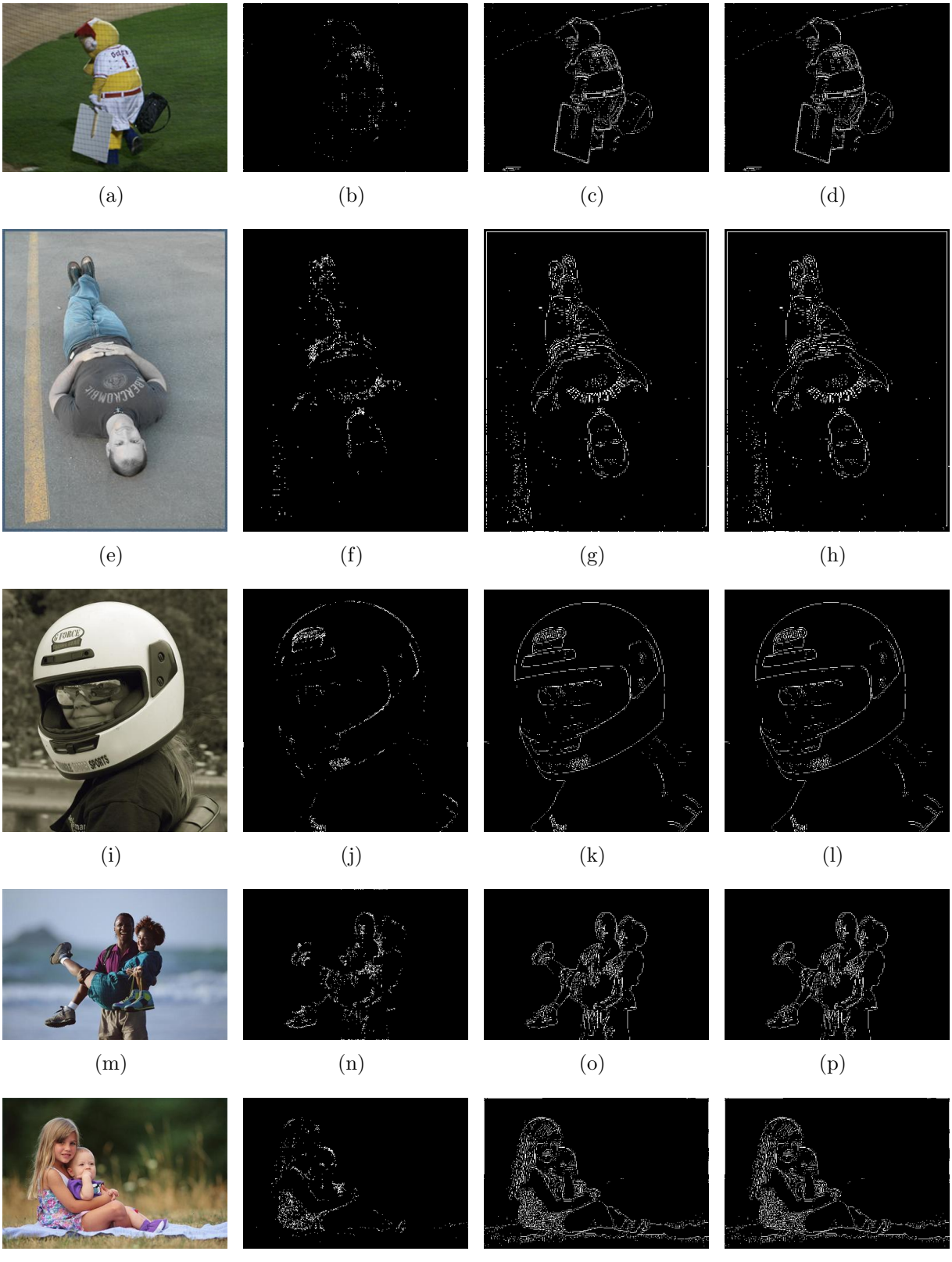
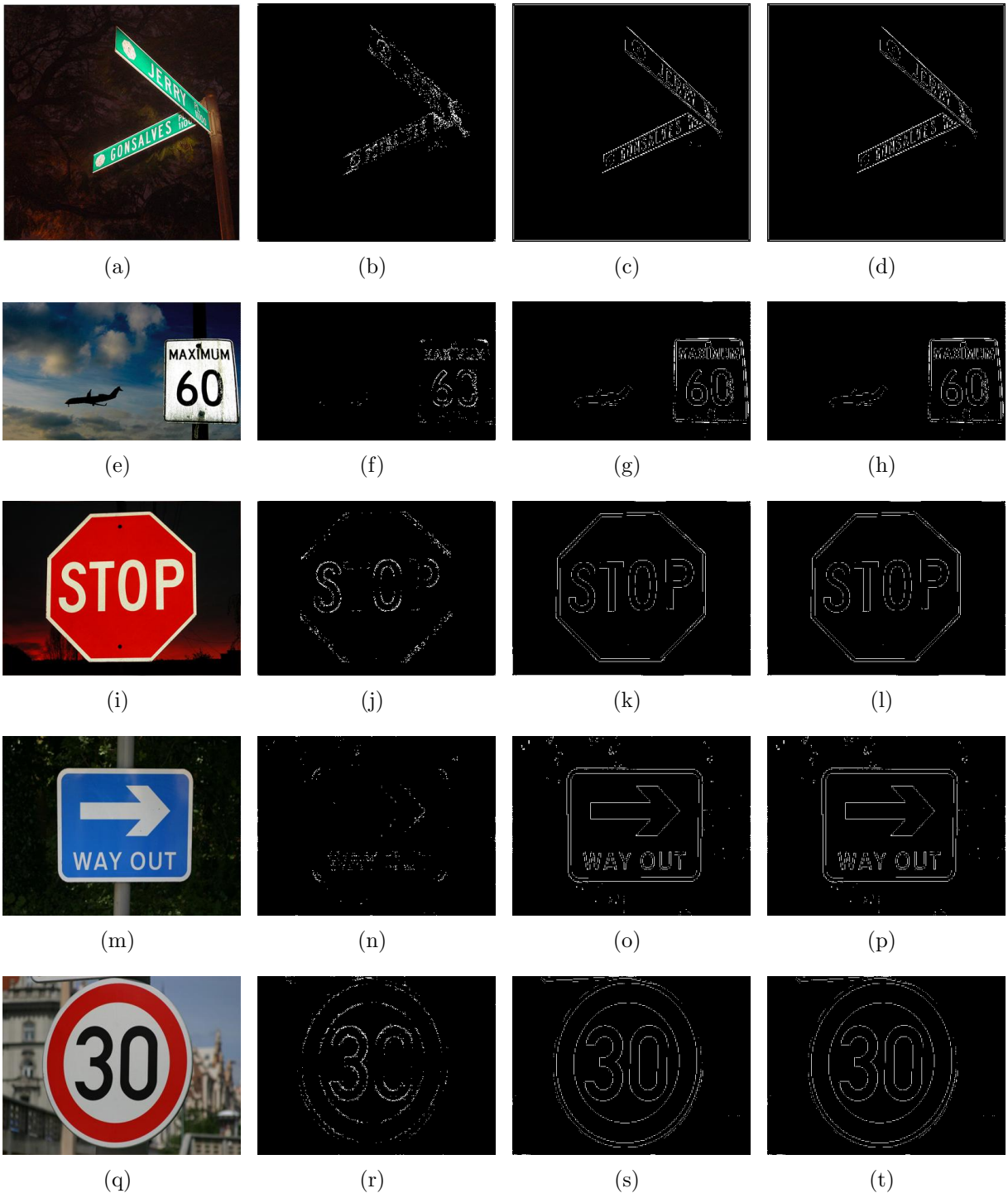


Figure 4.11: Comparison of human image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

Signs are perhaps the best example of the WR's and FTWR's ability to detect edges. This is most likely due to the nature of signs, and the high frequencies they possess. Figure 4.12 shows the results of some sign images and their respective salient edge maps. As is evident in Figures 4.12(f), 4.12(j), and 4.12(r), the SR method tends to highlight the numbers and letters of the sign, and suppress the borders. Both the WR and FTWR tend to pick up both the numbers and letters, as well as the borders; this is evident in all of the results in Figure 4.12. It should be noted however that in Figures 4.12(g) and 4.12(h), the airplane in the background is also falsely picked up by the WR and FTWR respectively, but is correctly suppressed by the SR method.



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

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Figure 4.12: Comparison of sign image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

Figure 4.13 shows a set of examples that contain images of animals and the results yielded from the salient edge detectors in this thesis. Like the sign images in Figure 4.12, the frequencies present in images of animals allow all three salient edge detectors to work well. It can be noticed that in almost all of the results, very little noise is picked up. This is most likely due to the consistently smooth nature of the background in the images. The WR and FTWR again outperform the SR method when it comes to detecting an object in it's entirety, as is evident in every example in Figure 4.13.

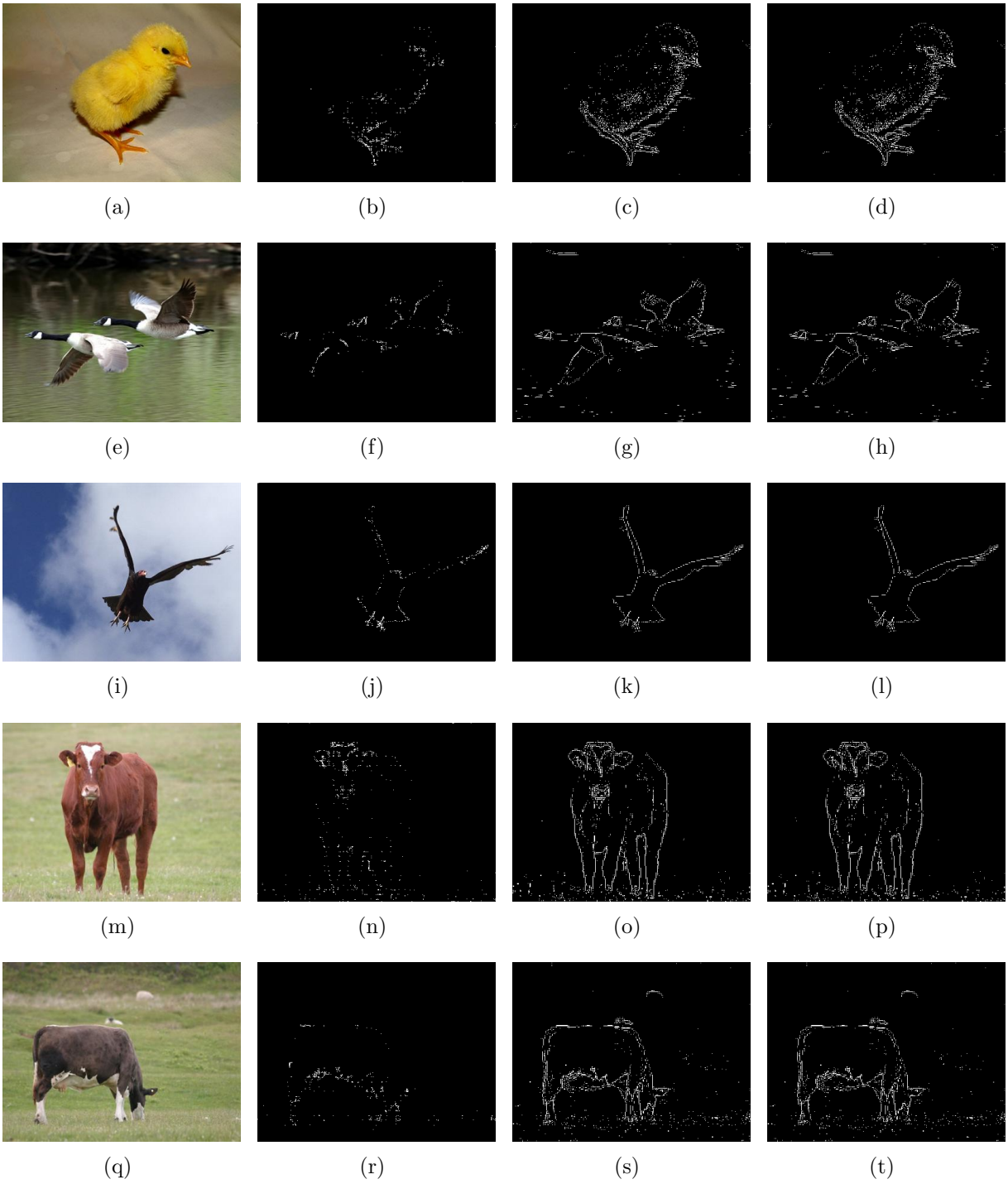


Figure 4.13: Comparison of animal image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

Figure 4.14 is a good example of the attention to detail that the WR and FTWR possess.

In the results yielded from Figure 4.14(a), the SR method only picks up the middle pepper (Figure 4.14(b)), whereas the WR and FTWR pick up all three peppers (Figures 4.14(c) and 4.14(d) respectively). In the results yielded from Figure 4.14(f), the saliency map produced by the SR method (Figure 4.14(f)) looks like pure noise, whereas the WR (Figure 4.14(g)) and FTWR (Figure 4.14(h)) pick up the entire screen and some of the details of what is being displayed on the screen. The same applies to the remainder of the examples in Figure 4.14, especially the results yielded from Figure 4.14(m), where the SR method does not clearly outline any of the numbers on the clock, but the WR and FTWR do.

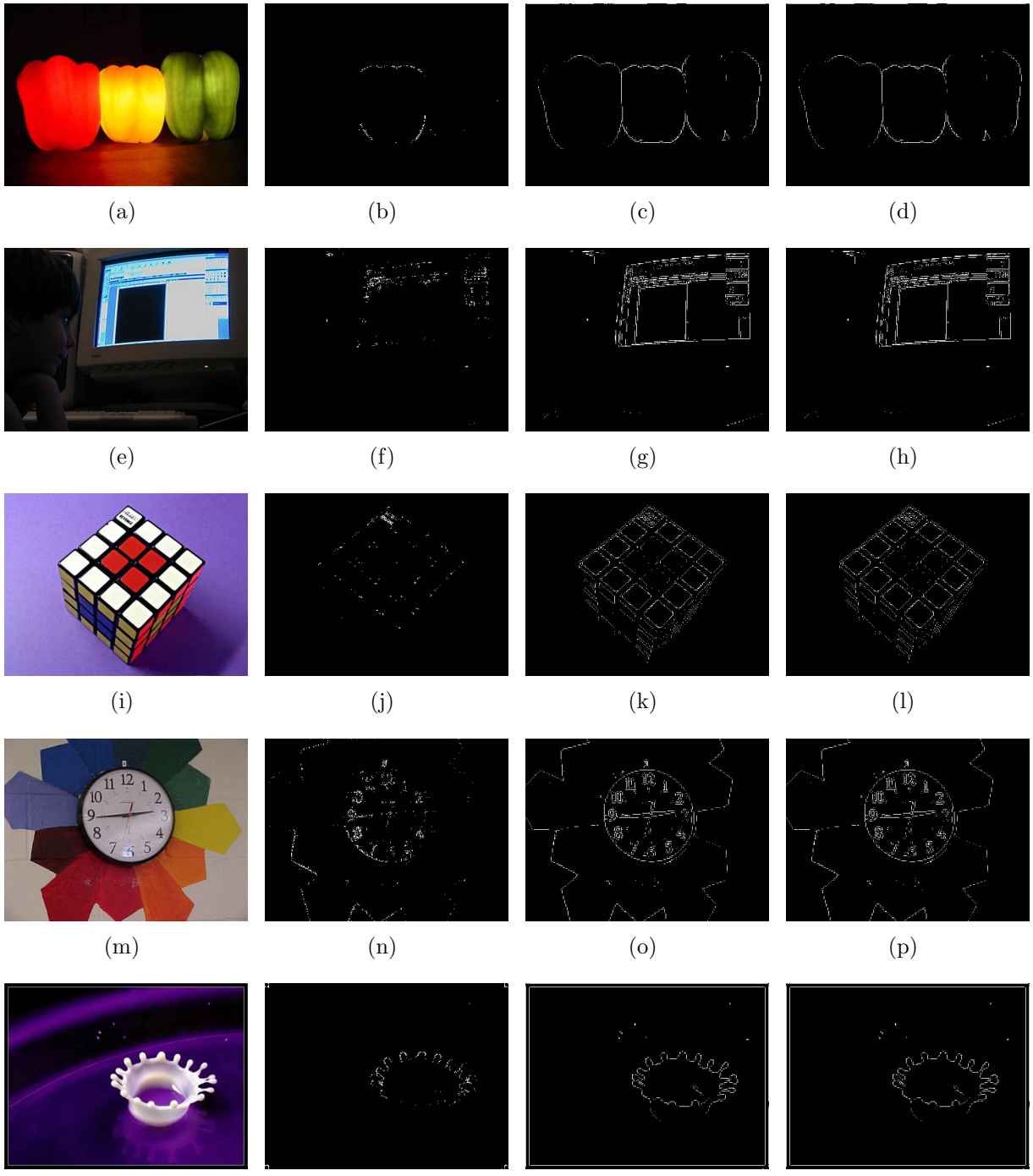


Figure 4.14: Comparison of miscellaneous image results using salient edge detectors. From left to right: Original image, SR's results, WR's results, and FTWR's results.

4.3 Summary

In this section, results of each method specified in this thesis are shown and compared with their respective competitors. For the salient region detectors, the proposed method (HDHVS) is compared with Itti's method, Achanta's method, and the spectral residual method. For the salient edge detectors, the proposed methods (WR and FTWR) are compared with the modified spectral residual method.

Over 5000 images were tested in this process (for each method), and a solid representation of the strengths and weaknesses of the proposed methods were presented in this section. To group the test images (for a better comparison), five categories of images were created: flowers, cars, human figures, signs, animals, and miscellaneous. This categorization allows the reader to compare images that are similar in nature (for example, all human figure images contain some skin component, and all sign images contain metal or wood).

Chapter 5

Conclusions and Future Work

5.1 Conclusions

IN this thesis, the issue of salient region and edge detection in images is addressed, extending the works of Itti et al [4], Achanta et al [6], and Hou et al [5] by introducing wavelets as a lossless resizing tool.

Itti's work uses Gaussian pyramids to downsample images to various scales. This introduced a significant loss in resolution and forces the final saliency map to be very blurry, displaying a blob of the salient object, rather than an exact outline and highlight of it. By using wavelets rather than Gaussian pyramids to resize the images, the loss of resolution is no longer experienced.

The spectral residual (originated by Hou) finds saliency by converting an image to the frequency domain and reconstructing it using only the frequency information. The same process is applied in this thesis except that rather using the Fourier transform, the discrete wavelet transform (DWT) is utilized. Using the DWT in this scenario allows for a segregation of the image's bandwidth. This allows certain frequencies to be picked up, while others are disregarded.

Achanta's work utilizes a frequency-tuning operation that auto-correlates the image to find the most prominent pixels in the image. Although no improvements are suggested for Achanta's method, the frequency-tuning operation is taken and applied to the detail components of a wavelet decomposition to produce a salient edge detector.

The concept of a salient edge detector has never before been mentioned in the literature. As the pioneering work in the field, the foundation is laid out for future work to improve and build upon. Along with suggesting two novel approaches to salient edge detection, a modification to the SR is also suggested, implemented, and used for comparison. Salient edge detectors can be a very useful tool for things like image segmentation or surveillance. Processing time is always an important factor regardless of the task at hand. This makes salient object detection (be it regions or edges) all the more important and applicable. When processing is applied only to salient regions, processing time is cut down drastically, depending on the size of the salient object.

It is discovered from this work that biologically inspired methods can coexist with computational efficiency. This is at least the case with the HDHVS method. Also, the importance of feature selection is displayed. After the establishment of Itti's method, most of the proceeding biologically inspired methods worked strictly with intensity, colour, and orientation. In this thesis, it is shown that good results can be obtained without using orientation which, in general, is the bottleneck to the processing speed of an algorithm.

5.1.1 List of Thesis Contributions

1. Inspired by Itti's approach, this thesis proposes the use of Gaussian pyramids with wavelet decomposition/reconstruction for lossless resizing. This allowed for high-definition saliency map construction, which is the basis of the newly proposed HDHVS method. This work has resulted in the publication: "High Resolution Biologically Inspired Salient Region Detection", accepted at the 2011 IEEE International Conference on Image Processing.

2. Inspired by the spectral residual method, a new method was proposed for constructing a salient edge detector. Using the same concepts (decomposing an image into the frequency domain and reconstructing using only phase information), the Fourier transform was replaced with wavelet decomposition to allow for a separation of the frequency spectrum. This is the basis of the proposed wavelet residual (WR) method.

3. Using both the spectral residual method as well as the frequency tuned method, another salient edge detector was proposed in this thesis. Along with the concept of decomposing an image into the frequency domain and reconstructing using only the phase information, the frequency tuning operator was also applied to tweak the salient edges and make them more vivid. This is the basis of the proposed frequency tuned wavelet residual (FTWR) method. This work resulted in the publication: "Frequency Tuned Salient Edge Detection", accepted at the 2011 IEEE Canadian Conference on Electrical and Computer Engineering.

5.2 Future Work

5.2.1 Time-Frequency Transforms

Considering that the novel advancements in this thesis revolve around wavelets, it is intuitive and natural to assume that the next step would be to implement the same conceptual ideas but using a more powerful time-frequency transform than wavelets. The Fourier transform was the first solid time/space to frequency transform and although it was very beneficial, it was not without it's limitations and drawbacks. The Fourier transform allowed a conversion to the frequency domain, but did not give any indication regarding the spatial location of frequencies. The DWT overcame this issue but is also not without its drawbacks. The continuation of time/space to frequency transform technology would only improve the already prevalent saliency detection methods that use this tool. This is because, if the one of the major tools used is improved, then the overall algorithm also improves.

5.2.2 The HDHVS Method

The HDHVS method can potentially be advanced by improving the computational efficiency of the centre-surround mechanism. If the same result can be produced without requiring another scale of the image, processing time would potentially decrease two fold.

Rather than limiting colour opponency to red-green and blue-yellow, the entire spectrum of the LUV colour circle could be utilized to allow for both truer opponencies as well as

Figure 5.1: LUV colour circle.

quantitatively more opponencies. As is seen in Figure 5.1, blue and yellow are not direct opposites. Rather, blue-orange and yellow-violet are true opposites.

Constructing a formula to extract the exact colour of a pixel and comparing it with its polar opposite across the entire LUV colour circle would potentially drastically improve the power of the colour opponency component of the HDHVS method.

Although a feasible method to model orientation for the HDHVS method was not discussed in this thesis, it does not exclude orientation as an important feature for saliency detection (for both edges and regions). Finding a suitable model for orientation that works with the HDHVS would improve the results. On this same note, modeling other features such as hue or texture could also potentially improve the results although it should be kept in mind that when it comes to feature selection, too many might actually hurt the final results rather than improve them, as is mentioned in [32].

5.2.3 The WR and FTWR Methods

Regarding the WR, no foreseeable improvements can be thought of besides applying the same ideology to a better time-frequency transform.

Regarding the FTWR, not only would a better time-frequency transform improve results, but a better auto-correlation function would also improve the results. The FTWR uses the frequency-tuning operator which simply takes the difference between each pixel value and the average pixel value from the entire image. Perhaps performing the same operation but only on a smaller portion of the image would improve the accuracy of the operator. Perhaps performing the operator on a moving window across the entire image would also improve results, and this would beg such questions as whether to overlap the window or not, how many iterations would run, the size of the window, and whether or not the size of the window would change with each iteration.

Appendix A

Thesis Related Publications

1. Yusuf Saber and Matthew Kyan, "Frequency Tuned Salient Edge Detection", Accepted in IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2011.

2. Yusuf Saber and Matthew Kyan, "High Resolution Biologically Inspired Salient Region Detection", Submitted to IEEE International Conference on Image Processing (ICIP), 2011.

Appendix B

Gaussian Blur

This section is in reference to Achanta's frequency tuning method. It is provided for the sole purpose of providing the reader with auxiliary information on said method, in case it is desired. The following is largely in reference to [6].

Let ω_{lc} be the low frequency cut-off value and ω_{hc} be the high frequency cut-off value. Only low frequencies in the original image must be considered since large salient objects are targeted. Thus, ω_{lc} must be fairly low; this satisfies the first criterion. This aspect also allows for a uniform highlighting of the salient object, which is the demand of the second criterion. High frequencies from the original image must also be retained in order to satisfy the third criterion: having well defined boundaries. Thus, ω_{hc} should be relatively high. However, the absolute highest frequencies must be suppressed in order to avoid noisy artifacts; this is the requirement of the fourth criterion. Since a saliency map with a vast array of frequencies is desired, combining several band pass filters with contiguous $[\omega_{lc}, \omega_{hc}]$ pass bands is appropriate.

For bandpass filtering, the difference of Gaussians (DoG) filter is used (see Equation B.1). As a side note, the Laplacian of Gaussian (LoG) is widely cited as the most satisfactory operator for detecting intensity changes when the standard deviations of the Gaussians are in the ratio 1:1.6. The DoG is a very good approximation of the Log, and this is why it is used. The DoG filter is given by:

$$DoG(x,y) = \frac{1}{2\pi}\left[\frac{1}{\sigma_1^2}e^{-\frac{(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2^2}e^{-\frac{(x^2+y^2)}{\sigma_2^2}}\right] \tag{B.1}$$

where σ_1 and σ_2 are the standard deviations of the Gaussian ($\sigma_1 > \sigma_2$).

The DoG's band is controlled by the ratio $\sigma_1 : \sigma_2$.

Continuing on to define $\sigma_1 = \rho\sigma$ and $\sigma_2 = \sigma$ so that $\rho = \sigma_1/\sigma_2$, we find that a summation over DoG with standard deviations in the ratio ρ results in:

$$\sum_{n=0}^{N-1} G(x,y,\rho^{n+1}\sigma) - G(x,y,\rho^n\sigma) = G(x,y,\rho\sigma^N) - G(x,y,\sigma)$$

for an integer $N \geq 0$, which is the difference of two Gaussians whose standard deviations are allowed to have any ratio $K = \rho^N$. Therefore, by choosing a DoG with a large K, the combined result of applying several band pass filters can be obtained. Assuming that σ_1 and σ_2 very in such a way that ρ remains at a value of 1.6, then, what is essentially happening is the outputs of several edge detectors are being added, at several image scales.

A concise selection of σ_1 and σ_2 should yield an appropriate bandpass filter that retains spatial frequencies from the original image. If σ_1 and σ_2 have a great difference between them, the pass band of the resulting band-pass filter given in Eq. B.1 can be approximated from the two constituent Gaussians. With $\sigma_1 > \sigma_2$, ω_{lc} is determined by σ_1 and ω_{hc} is determined by σ_2. Unfortunately, using these filters at a practical length, providing a correspondingly simple implementation, makes this approximation rather inaccurate.

Both σ_1 and σ_2 are selected as follows: for a large ratio in standard deviations, σ_1 is set to infinity. This results in a notch in frequency at DC while retaining all other frequencies.

To remove high frequency noise and artifacts, a small Gaussian kernel is used, keeping in mind the need for computational simplicity. For smaller kernels, the binomial filter approximates the Gaussian well. $\frac{1}{16}$[1,4,6,4,1] is used, giving $\omega_{hc} = \pi/2.75$. Therefore, more than twice as much high-frequency content is kept from the original image as GB and at least 40% more than SR.

Bibliography

[1] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, Vol. 2, No. 3, pp. 194-203, 2001.

[2] O. Spakov and D. Miniotas, "Visualization of eye gaze data using heat maps," *Mendeley Electrical Engineering*, Vol. 2. Issue 2, pp. 55-58, 2007.

[3] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini, "Adaptive color image compression based on visual attention," *11th IEEE International Conference on Image Analysis and Processing*, 2001.

[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, No. 11, 1998.

[5] L. Z. X. Hou, "Saliency detection: A spectral residual approach," *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[7] W. James, "The principles of psychology," *Harvard University Press, Cambridge, Massachussetts*, 1980.

[8] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of Vision*, Vol. 8. Issue 3, pp. 1-15, 2007.

[9] L. Itti and C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems," *SPIE Human Vision and Electronic Imaging IV*, pp. 473482, May 1999.

[10] H. X. Q. Zhang, "Extracting regions of interest in biomedical images," *International Seminar on Future BioMedical Information Engineering*, 2008.

[11] R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," *IEEE 17th International Conference on Image Processing*, 2010.

[12] M. Saad and A. Bovik, "Extracting regions of interest from still images: Color saliency and wavelet-based approaches," *5th IEEE Signal Processing Education Workshop on Digital Signal Processing and DSP/SPE*, 2009.

[13] Y. Ma and H. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *Proceedings Of The Eleventh ACM International Conference On Multimedia*, 2003.

[14] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, pp. 545552, 2007.

[15] T.Ohashi, Z.Aghbari, and A.Makinouchi, "Hill-climbing algorithm for effcient color-based image segmentation," *International Conference On Signal Processing, Pattern Recognition, and Applications*, June 2003.

[16] N. Bruce and J. Tsotsos, "Attention based on information maximization," *Journal of Vision*, Vol. 7, No. 9, pp. 950956, 2007.

[17] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," *International Conference on Computer Vision Systems*, March 2007.

[18] Q. Zhang, Y. Zheng, and H. Xiao, "Automatically extracting salient regions in natural images," *International Colloquium on Computing, Communication, Control, and Management*, 2009.

[19] Y. Hu, X. Xie, W. Ma, L. Chia, and D. Rajan, "Salient region detection using weighted feature maps based on the human visual attention model," *Springer Lecture Notes in Computer Science*, Vol. 3332, No. 2, pp. 9931000, October 2004.

[20] Q. Zhao, Y. Hu, and J. Cao, "Automatic image segmentation based on saliency maps and fuzzy svm," *IET International Communication Conference on Wireless Mobile and Computing)*, 2009.

[21] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation," *International Conference on Computer Vision Systems*, Vol. 5008, pp. 6675, 2008.

[22] T. N. Mundhenk, C. Ackerman, D. Chung, N. Dhavale, B. Hudson, R. Hirata, E. Pichon, Z. Shi, A. Tsui, and L. Itti, "Low cost, high performance robot design utilizing off-the-shelf parts and the beowulf concept, the beobot project." *Proc. SPIE Conference on Intelligent Robots and Computer Vision XXI: Algorithms, Techniques, and Active Vision*, pp. 293-303, 2003.

[23] R. Wildes, "A measure of motion salience for surveillance applications," *International Conference on Image Processing*, 1998.

[24] Y. Xia, Y. Gan, W. Li, and S. Ning, "A simple and fast region of interest extraction approach based on computer vision for sport scene images," *2nd International Congress on Image and Signal Processing*, 2009.

[25] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, Vol. 12, No. 1, 1980.

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

[27] E. Gelasca, D. Tomasic, and T. Ebrahimi, "Which colors best catch your eyes: A

subjective study of color saliency," *Fisrt International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005.

[28] J. Daugman, "Complete discrete 2d gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36. No. 7, pp. 11691179, 1988.

[29] 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*, 2008.

[30] C. Ngau, L. Ang, and K. Seng, "Bottom-up visual saliency map using wavelet transform domain," *3rd IEEE International Conference on Computer Science and Information Technology*, 2010.

[31] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2007.

[32] R. Duda, P. Hart, and D. Stork, *Pattern Classification, 2nd Edition.* Wiley-Interscience, 2000.

[33] N. Butko, L. Zhang, G. Cottrell, and J. Movellan, "Visual saliency model for robot cameras," *IEEE International Conference on Robotics and Automation*, 2008.

[34] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Transactions On Multimedia*, Vol. 11, No. 5, 2009.

[35] W. Cheng, W. Chu, J. Kuo, and J. Wu, "Automatic video region-of-interest determination based on user attention model," *IEEE International Symposium on Circuits and Systems*, 2005.

[36] L. Wei, L. Yong, R. Yi, and D. Peng, "A new approach for extracting the contour of an roi in medical images," *International Conference on Advanced Computer Theory and Engineering*, 2008.

[37] C. Ngau, L. Ang, and K. Seng, "Comparison of colour spaces for visual saliency," *International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2009.

[38] G. Hu, M. Xiao, and S. Yuan, "Detecting automatically and compression algorithm for infrared image based on region of interest," *International Forum on Computer Science-Technology and Applications*, 2009.

[39] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, 10(1), 2001.

[40] H. S. N. Vasconcelos, "Object-based regions of interest for image compression," *Data Compression Conference*, 2008.

[41] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, 40 (10-12). pp. 1489-1506, 2000.

[42] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," *IEEE International Conference on Computer Vision*, pp. 16, Oct. 2007.

[43] R. Achanta and S. Susstrunk, "Saliency detection for content aware image resizing," *IEEE International Conference on Image Processing*, 2009.

[44] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *Journal of Optical Society of America*, Vol. 23, No. 10, pp. 24622470, October 2006.

[45] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*, pp. 59-68, October 2005.