Wavelet-based Scale Saliency

Anh Cat Le Ngo\textsuperscript{a,b}, Kenneth Li-Minn Ang\textsuperscript{c}, Guoping Qiu\textsuperscript{b}, Jasmine Kah Phooi Seng\textsuperscript{d}

\textsuperscript{a}School of Electronic Electrical Engineering, The University of Nottingham, Malaysia Campus
\textsuperscript{b}School of Computer Science, The University of Nottingham, UK Campus
\textsuperscript{c}Centre for Communications Engineering Research, Edith Cowan University
\textsuperscript{d}Department of Computer Science & Networked System, Sunway University

Abstract
Both pixel-based scale saliency (PSS) and basis project methods focus on multiscale analysis of data content and structure. Their theoretical relations and practical combination are previously discussed. However, no models have ever been proposed for calculating scale saliency on basis-projected descriptors since then. This paper extend those ideas into mathematical models and implement them in the wavelet-based scale saliency (WSS). While PSS uses pixel-value descriptors, WSS treats wavelet sub-bands as basis descriptors. The paper discusses different wavelet descriptors: discrete wavelet transform (DWT), wavelet packet transform (DWPT), quaternion wavelet transform (QWT) and best basis quaternion wavelet packet transform (QWPTBB). WSS saliency maps of different descriptors are generated and compared against other saliency methods by both quantitative and quantitive methods. Quantitative results, ROC curves, AUC values and NSS values are collected from simulations on Bruce and Kootstra image databases with human eye-tracking data as ground-truth. Furthermore, qualitative visual results of saliency maps are analyzed and compared against each other as well as eye-tracking data inclusive in the databases.

Keywords: visual attention, visual saliency, scale saliency, discrete wavelet transform, quaternion wavelet transform, wavelet packet best basis

1. Introduction
A few centuries ago, Neisser proposed a fundamental theory about the human visual attention system including pre-attentive and attentive stages...
in his psychology studies. However, his work was unknown to machine vision scientists until David Marr, a neurologist, proposed a neurology-based computational model for Neisser’s theory. The computational model includes a feature extraction stage followed by perceptual grouping stage. Though the model was practically limited and rarely implemented, it inspires and provides framework for several later computational models. Among them, Itti model holds significant influence and provides a standard in the research field. Itti models feature extraction as center-surrounds operator, a property of visual cortex; while, perceptual grouping and attentive region assessment are due to Proto-Object generation and Winner-Take-All network. After center-surrounds operations in multi-scale levels were proposed for construction of conspicuity and saliency maps by Itti et al, other theories like Graph-based Visual Saliency, and Spectral Residual Saliency were brought in to produce more meaningful saliency maps as well as reduce the computational complexity. These saliency models assume that human vision systems may behave like random-walk processes or follow statistical property of natural images.

Without making such strong assumption, Kadir and Gilles initiated information-based saliency map with their work on pixel-based scale saliency (PSS). Other information-related saliency research rapidly gained pace with Niel Bruce’s An Information Maximization (AIM) theory and Danash Gao’s Discriminative Information Saliency (DIS). Furthermore, the information-based spatial-temporal framework (ENT) extends and fastens the models from still images to the dynamic video context.

Information-based saliency approaches all motivated from the assumption that human attention could be attracted to spatial location accompanied with highly informative content. From signal coding, compression and self-information theory, an event has more information when it appears to be structural and rare. Though based on similar concepts, each method has its own information estimation approach on different type of descriptors. Generally, those approaches can be characterized according to their choices of descriptors and calculation methods. For examples, PSS and ENT utilize pixel-value descriptors; meanwhile, AIM and DIS emphasizes on the alternative basis-projection descriptors, ICA bases and Wavelet bases consecutively. In accordance with information measurement, ENT and PSS employ the popular Shannon entropy estimated by histogram construction or Parzen kernel. AIM estimates self-information through neural-network training on patches of natural images. Decision-theory based DIS has its
discriminative information from classifying descriptors into center or surrounds classes. Noteworthy, PSS is so far the only approach accumulating information of both descriptors and their structure. However, PSS employs pixel-value descriptors and isotropic circular sampling, which might hinder its performance in term of accuracy due to failure in extracting popular oriented features in natural images as well as speed due to the curse of dimensionality in information estimation.

The limitation of PSS sparked discussion for alternative solutions by Kadir et al. [7] [13]. Deploying basis-projected descriptors in place of pixel-based descriptors not only boosts practical performance of scale saliency but as well provides deeper theoretical understanding of scale saliency and data multi-scale structural information. Moreover, the extension would make scale saliency the first-ever method capable of using both pixel-value descriptors (PSS) and basis-projection descriptors. Wavelet elements are preferred as alternative basis in this paper; therefore, the proposed method is named Wavelet Scale Saliency (WSS). In order to clearly explain extension from pixel-based descriptors to basis-projection descriptor, we organize this paper in following sections. Section 2 gives overview about scale saliency and its main idea. The next section 3 explains the rationale behind usage of time-frequency domain instead of time-domain only for visual saliency; meanwhile, sections 4, 6 elaborates statistical distribution and correlation of time-frequency descriptors. As wavelet is chosen as time-frequency basis, section 5 gives background information about four types of wavelet transforms considered in this study: discrete wavelet transform (DWT), discrete best basis wavelet packet transform (DWPTBB) as well as two quaternion wavelet transforms QWT and QWPTBB. Accordingly, there are four time-frequency descriptors representing time-frequency domain slightly different from each other. Moreover, each descriptor depends on a particular morphological shape of its own mother wavelet. All details about properties of those descriptors are organized in section 6.1. Along with new descriptors, suitable mathematical models of feature-space and inter-scale saliency estimation are derived in section 7. Moreover, the mathematical derivation unveils strong relation between WSS with another state-of-the-art Bayesian Surprise Saliency (BSS) [14]. Beside theoretical evaluation, simulations on Neil Bruce image database [15] and Kootstra image database [16] are carried out in order to compare quantitatively the proposed WSSs with different basis-projection descriptors, the original PSS and the ITT model. Furthermore, qualitative analysis on particular images provides better details about
responses of the proposed methods with different types of scenes. It is possible that performance of saliency methods depends on image content. All results and discussion are detailed in the section [8]. Finally, the conclusion [9] summarizes our main contributions and future research directions.

2. Scale Saliency

To get a hold of what exactly is a scale saliency; a few fundamental principles of original scales saliency are reviewed. Scale saliency utilizes maximum feature-space entropy weighted by its inter-scale dependency across scales as saliency values; furthermore, it argues that information measurement might be data-driven pivot for human visual attention. Its mathematical model is summarized as follows.

\[
Y_D(s_p, \vec{x}) = H_D(s_p, \vec{x})W_D(s_p, \vec{x}) \tag{1}
\]

\[
H_D(s_p, \vec{x}) = -\int_{d \in D} p(d, s_p, \vec{x}) \log_2 (p(d, s_p, \vec{x})) \, dd \tag{2}
\]

\[
W_D(s_p, \vec{x}) = s \int_{d \in D} \left| \frac{\delta p(d, s_p, \vec{x})}{\delta s} \right| \, dd \tag{3}
\]

\[
s_p = \{ s \mid \frac{\delta H_D(s, \vec{x})}{\delta s} = 0; \frac{\delta^2 H_D(s, \vec{x})}{\delta s^2} < 0 \} \tag{4}
\]

Feature-space saliency, \((H_D)\) in the equation [2] is measured by its Shannon entropy of pixel-values descriptor \((d)\) at a specific scale or sampling window size \((s)\) for each image location \((\vec{x})\). Shannon entropy is chosen since it satisfies fours over five criteria of multi-scale entropy filtering [17]. The last criterion actually requires structural correlation from information estimation, which is apparently not considered in Shannon entropy. However, inter-scale saliency, \((W_D)\), actually fulfils this requirement, and it is estimated at every location by total variation of descriptors’ probability distribution function \((PDF)\) across scales. Then, the scale \((s_p)\) at which most significant information should be found; it is actually the maximum point of the scale-entropy concave curve in the equation [4]. Finally, the overall saliency is stated mathematically as the equation [4] in accordance with the definition of scale saliency. Let’s apply the concept of scale saliency on a general form of signal \(R(x_0, s_i) = \{ I(x_0, s_i) + N(x_0, s_i) | i = 1...n \}\), where \(I_{x_0, s_i}\) is ideal noise free signal, \(R_{x_0, s_i}\) is the measured signal with noise \(N_{x_0, s_i}\) at specific location and scale \((x_0, s_i)\). Assumed no dependencies between noise and the
ideal signal, the estimated entropy is $H_D(\vec{R}_{\vec{x}_0,s_i}) = H_D(\vec{I}_{\vec{x}_0,s_i}) + H_D(\vec{N}_{\vec{x}_0,s_i})$.

Assumed that noise PDF are scale-invariant, the equation 3 implies that inter-scale saliency measure is purely dependent on variation of useful signal $\Delta_s H_D(\vec{I}_{\vec{x}_0,s_i})$ and not affected by variation of noise $\Delta_s H_D(\vec{N}_{\vec{x}_0,s_i}) = 0$.

This briefly explains basic motivation behind scale saliency work; further mathematical analysis and experiments results can be found in [7] [13].

The original scale saliency [7] uses pixel-value descriptors which are simple, intuitive, and straightforward interpretation of image data. Moreover, its combination with circular sampling window provides isotropic information analysis, independent of any morphological shape inside sampled regions. Nevertheless, its drawbacks are susceptible to noise, require high computational cost and cause significant bias in entropy estimation. So far, histogram construction and approximated Parzel kernel are two popular parameter methods for constructing pixel-value descriptors’ PDF and estimating entropy. Entropy bias and speed performance in those mentioned methods greatly depend on manual tuning of histogram numbers of bins or Parzel size kernel; in addition, they as well restrict extension of scale saliency to higher dimensional data. Suau [18] overcomes these problems by bypassing pdf construction stage and estimating PSS by multivariate-data-adaptive information estimation technique [19]. In spite of its fast computation for multivariate data, the non-pdf approach hinders the inter-scale saliency process which directly depends on PDF’s 3. It is solved by adapting set-theory based elegant solutions of Kadir [13] for inter-scale saliency $W_D$ computation into kd-tree structure. However, the solution is not intuitively and mathematically related to the information-based framework. That motivates us develop (WSS), a more coherent information-based scale saliency with subband energy descriptors, as solutions for all these short-comings of PSS.

3. Time-Scale-Frequency

A well-known computational model of visual attention is first mentioned in Koch and Ullman’s publication [20]. After that, several other models are proposed; however, they are usually over-complex and not biologically plausible. The disadvantages might be due to pixel representation utilized in many early visual attention algorithm. To overcome these problems systematically, Urban [21] has investigated strong constraints to keep computational complexity within an acceptable range for possible real-time implementation. These constrains are drawn from evidences of psychological experiments.
which shows that images could be analyzed in psycho-visual channels at least in TV-viewing condition [22]. In other words, visual data could be further analysed into channels and sub-bands instead of being used in raw pixel format. Furthermore, the channels can be effectively characterized by separated frequency bands and orientation ranges of wavelet analysis [23].

Let's assume visual active areas of brains can deploy some 9/7 Cohen-Daubechies-Feauveau (CDF) wavelet transform operators; then, it results in multi-scale pyramid composed of oriented contrast maps with limited frequency range and low-resolution image. For each level of wavelet decomposition, there are four channels: (i) sub-band 0 is approximated image after filtered with many low-pass blurring kernels; (ii) sub-band 1 extracts horizontal frequencies corresponding to vertical edges of images (iii) sub-band 2 contains frequencies and features along two diagonals of image frames. (iv) sub-band 3 prefers vertical frequencies mapping to horizontal features form images. Natural scenes are full of horizontal, vertical or two diagonals features; therefore, human visual perception seems to prefer those dominant features. Besides oriental constraints, visual acuity is another visually perceptual limit. Normally, human fovea could decompose and process details above its limit visual acuity (1.5-2 degrees of visual angle). It lasts in frequency range: 0.7-0.5 pixels per degree, or 0.33-0.25 cycles per degree. This range is nearly resembled by the last level low-resolution version of images in usual wavelet decomposition. Each decomposed level are generated by moving kernels with different window size to any image positions. Spatial frequency of other wavelet analysis levels, varying in accordance with analyzing depths, is shown the following table [1]

Spectral energy are usually employed as spectral signature for image collections or individual images. Urban et al [21] analyses different sets of images belonging to four different semantic categories: coast, mountain, street and open-country. Interestingly, Fourier spectrum of each category possess distinguished shape and frequency range, significantly different from each other [21]. In other words, each general spectral profile and associated distribution histogram of image classes have unique energy distribution. This distribution

| Depth | 0   | 1   | 2   | 3   | 4   | 5   |
|-------|-----|-----|-----|-----|-----|-----|
| Frequency range | 10.7-5.3 | 5.3-2.7 | 2.7-1.3 | 1.3-0.7 | 0.7-0.3 | 0.3-0.2 |

Table 1: Wavelet Levels vs Frequency Range
is proportional to distance $d$ from mean magnitude spectrum normalized by the standard deviation of the category.

$$d(s) = \frac{1}{nm} \sum_{n,m} |s(n,m) - AS(n,m)|$$

where $AS(n,m)$ is average spectrum. Carefully observing the spectral profiles could give distinguishing clues for each semantic scene. For example, "Coastal" scenes are dominated with horizontal features; therefore, its spectral profiles stretch along vertical axes. Furthermore, spectral profile of "OpenCountry" categories is biased toward two upper and lower spectrum. Though almost similar to the "OpenCountry" profile, spectrum of "Street" images includes more types of features from artificial environment beside horizon-oriented details. Therefore, the diamond of image spectral profile of "Street" becomes more significant horizontally. "Mountain" categories with its random scenic details have isotropic spectrum while scenes of streets filled with artificial objects have spectrum stretched in both horizontal and vertical axis. From Urban's research, spectral energy distribution seems to be important clues for visually perceptual system of human beings.

Beside image classification, the spectral distribution signature is as well useful in visual attention and early visual process. Such energy distribution becomes differentiable clues for features across scales. Spectral profiles of image feature at a particular scale would help differentiate itself from directly upper and lower scale. Lets do an imaginary experiments with a single square input signal $x(t)$ defined as follows.

$$x(t) = \begin{cases} 
1 & t_1 \leq t \leq t_2 \\
0 & t < t_1 \lor t > t_2 
\end{cases}$$

If $x(t)$ is filtered by a kernel $F()$ with kernel size (1-D kernel width) $W = \Delta T$ and $W$ is much smaller than non-zero period of the given square signal $\Delta T \ll |t_2 - t_1|$, the response will be just two impulse function at $t_1, t_2$.

$$F(x(t)) = \begin{cases} 
1 & t = t_1 \land t = t_2 \\
0 & t \neq t_1 \lor t \neq t_2 
\end{cases}$$

For 2-D signal or image context, the above operation corresponds to a classic edge detection phenomena. Though edges and structures plays important
roles in visual perception, their information does not sufficiently represent
the whole natural scenes. Natural images are rich of other features like tex-
ture, flat regions, etc beside edges and corners. As mentioned before, image
features can be interpreted in terms of energy distribution. For example,
edges are the places of high energy concentration, homogeneous flat regions
do not contain much energy while textures, hybrid of edges and flat regions,
contains certain amount of energy. If only one window size is used in the
analysis, significant responses only come from features or objects which hap-
pen to fit into that window size. The other useful features with inappropriate
size in accordance with the filter could not be extracted. Therefore, a multi-
scale’s approach is extremely necessary in order to identify suitable sizes of
kernels or fuse features from different scales together. When mother wavelet
is chosen as filtering kernels, window size becomes equivalent to frequency
range in the table I and choosing adaptive frequency ranges is important
computation task for spatial feature extraction. Inspired by such fundamen-
tal query in computer vision, this paper tries to contribute a little insight
about how spectral density distribution can characterize features at each
scale and how the frequency range of processing can be appropriately cho-
osen for multi-scale feature representation. From the multi-scale features and
appropriate scale selection, we can develop computation saliency methods
capable of highlighting salient features across scales by using spectral energy
distribution.

4. Time-Scale-Energy

PSS estimates information from pixel values, time-domain descriptors by
constructing normalized histogram of pixel values as probability distribution.

\[ p_h(d) = \frac{n_d}{N} \]

where \( p_h \) is probability of descriptor, the ratio between number of pixels with
d descriptors and total image pixels N. Lets use square of pixel \( d^2 \) as weights

\[ p_e(d) = \frac{n_d * d^2}{\sum_D (n_d * d^2)} = \frac{\rho_x}{E_x} = \frac{\int_{-\infty}^{\infty} (x(t)\delta(x(t) - d))^2 dt}{\int_{-\infty}^{\infty} x(t)^2 dt} \]
Normalized weighted-histogram $p_e(d)$, of signal $x(t) \in L^2(R)$ can obviously be interpreted as $\rho_x$ energy density of descriptors $d$ in time domain. By the isometric property of the Fourier transform, the PDF of energy density distribution can be expressed in frequency domain as well.

$$p_e(d) = \frac{\rho_x}{E_x} = \frac{\int \hat{x}(f)^2 \delta(\hat{x}(f) - \hat{d}) df}{\int \hat{x}(f)^2 df}$$

Or in joint time and frequency domain.

$$p_e = \frac{\rho_x}{E_{\rho_x}} = \frac{\rho_x(t,f)}{\int \int \rho_x(t,f) dtdf}$$

where

$$\rho_x = \left| \int x(\tau) g_{*,f}^*(\tau) d\tau \right|^2$$

where $\rho_x$ is energy density in joint time-frequency representation. Pure time descriptors have perfect localization in time, no localization in frequency, and vice versa for frequency descriptors. Both extreme time or frequency descriptors make interpretation of constructed PDF, and estimated information difficult to explain. Therefore, it is necessary to find a representation of $g_{*,f}^*(\tau)$ which describes spectral density of local energy. For example, Short-Time Fourier Transform (STFT) is the first-known transform capable of generating spectrogram, a graphical representation of local signal energy in time-frequency plan.

$$\rho_x(t,f) = \left| \int x(\tau) h^*(\tau - t) e^{-2j\pi f \tau} d\tau \right|^2$$

STFT identifies spectral density as well as local energy density or information in a short-time period of the signals. However, it does not much benefit scale saliency unless scale parameters are actually considered as in signal description on phase-space. Fortunately, in recent years, alternative scale-based representation, called wavelet-transform (WT), has been widely
addressed among signal processing community, and its fundamental idea is replacing the frequency shifting operation $e^{-2j\pi f\tau}$ by a time (or frequency) scaling operation $\psi(t-a\tau)$, a basic wavelet kernel. Consequently, the energy density in WT framework is formulated as follows.

$$\rho_x(t,a) = \left| \int_{-\infty}^{\infty} x(\tau) \psi(\frac{\tau - t}{a}) d\tau \right|^2$$

WT coefficients, $\rho_x(t,f)$, averagely measure spectral density of frequency sub-bands, a short range of frequency, in a short period of time. Characteristics of the time-frequency window are specified by two main parameters, time-shift $\tau$ and scale $a$. As derived from short-time spectral representation, a spectrogram, by utilizing scale operations, its energy density distribution is called scalogram. Given time-scale space, the total signal energy can be rewritten as follows.

$$E_{\rho_x} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho_x(t,a) dtda$$

and probability of time-scale descriptors can be specified with scale parameters

$$P_e(t,a) = \frac{\rho(t,a)}{E_{\rho_x}}$$

Generally, the wavelet transform can help generate frequency sub-band coefficients, square of which over total energy are density distribution of that sub-band in time-scale space.

5. Wavelet Transform

After discussing about usefulness of time-frequency-scale representation in the section 3 and its corresponding time-frequency energy distribution, we recognize that wavelet-representation would be ideal candidates for our investigation into energy density distribution and other statistical property across multiple scales of natural images. During the quite short history of wavelet analysis, this research fields have been very fruitful and there are several analysing techniques with wide range of characteristics. In this paper, only standard techniques such as discrete wavelet transform (DWT), discrete wavelet packet transform (DWPT), quaternion wavelet transform with
best-basis (QWTBB), and quaternion wavelet packet transform with best-basis. are deployed as possible descriptors. In this subsection, we first look into discrete and real wavelet and wavelet packet transform with best basis (DWT, DWPTBB) for its theoretical background; then QWT and QWPTBB, quaternion versions of two prior wavelet transforms, are considered.

5.1. Discrete Wavelet Transform

Though wavelets were firstly introduced in the early 20th century by Alfred Harr, they are only developed rapidly much later. Only until recently, they have been widely employed in many computer vision problems such as image or video de-noising, enhancement, coding, and pattern classification [24, 25, 26]. Signal analysis for frequency components can be achieved by Fourier transform (FT) but FT does not provide suitable tool for time-frequency analysis of images. Short-time Fourier Transform (STFT) is an extension from FT approach for analysing local frequency analysis at a short period of time [27]. Noteworthy that, STFT can be used for taking the spatial interval in 2-D signal instead of time period in 1-D type since there is no time dimension for still images. However, STFT utilizes fixed window kernels for every data blocks across input signals; this property make STFT less suitable for complex signal analysis, especially signals with strong semantic structures appearing across multiple scales. In other words, STFT only succeeds with signals whose features are embedded in fixed definite temporal or spatial regions or there is prior knowledge about a suitable size of window kernel for STFT processing. Without the above conditions, STFT would totally miss signal features. Theoretically, the disadvantage can be avoided by employing STFT with multiple kernel sizes; however, it raises up another issues such as what range of sizes would be chosen to optimally extract useful features with reasonably computational effort.

Problems of STFT in analyzing local frequency have motivated development of multi-scale wavelet techniques for better local frequency representation. Since limitations of STFT is due to fixed-size processing windows, wavelet analysis deploys multi-resolution filter-banks on input signals. As illustrated in figure 1, 1-D signals are decomposed into low-pass and high-pass components. In case of 2-D input signals, the filtered outputs are four sub-bands: low-low, high-low, low-high, and high-high in regards of processing orientation. Intuitively, 2-D signals analysis includes row-wise 1-D analysis followed by column-wise 1-D analysis or vice verse. With respect to processing direction, high-low sub-band tends to extract horizontal features, low-
high sub-band prefers vertical features, high-high sub-band detects diagonal features, and low-low are approximated version of original signal by inverse dyadic scale. Let’s assume that input signals are two-dimensional grey-scale image \( f(x, y) \), and the scaled mother wavelets have following mathematical form:

\[
\psi_{s,i}(x,y) = 2^s \psi(2^s x, 2^s y) \mid_{i=\{v,h,d\}}
\]

for vertical, horizontal and diagonal sub-bands and a scaling function \( \phi_S(x, y) = 2^S \phi(2^S x, 2^S y) \) for low-resolution signals with \( s \geq S \). Then, we can represent any images \( f(x, y) \in L_2(\mathbb{R}) \) as:

\[
f(x, y) = \sum_{x,y} c_S(x,y) \phi_S(x,y) + \sum_{i=\{v,h,d\}} \sum_{x,y,s \geq S} d_{s,i}(x,y) \psi_{s,i}(x,y)
\]

where

\[
c_S(x,y) = \int f(x,y) \phi_S(x,y) dx dy
\]

\[
d_{s,i}(x,y)\mid_{i=\{v,h,d\}} = \int f(x,y) \psi_{s,i}(x,y) dx dy
\]

\( c_S(x,y) \) and \( d_{s,i}(x,y) \) are scaling coefficients and wavelet coefficients from vertical, horizontal and diagonal sub-bands. The parameter \( S \) represents the lowest analyzing depth while \( s \) is higher decomposing levels in multiple scale-space framework. As mentioned before, 2-D DWT can be obtained by tensors products of 1-D DWT when the original image \( f(x, y) \) is analyzed along two dimensions \( x \) and \( y \) separately. As a result, the scale function \( \phi(x,y) \) is approximated as \( \phi(x)\phi(y) \) and filter-banks of three directional sub-bands are \( \phi(x)\psi(y), \psi(x)\phi(y), \psi(x), \psi(y) \). In the figure 1, discrete wavelet transforms are carried on the sample image in the left hand-side. On the right-hand side contains decomposed results by two levels with three distinctive sub-bands and a down-sampled version of the original image.

Noteworthy that, the real-wavelet transform like DWT suffers from shift-variance, a small shift in the signal can greatly change magnitudes of wavelet coefficients around singularities. Furthermore, it has no phase to embed signal location information therefore aliasing effects would be introduced into recovery process. These issues need seriously considering whenever the discrete real wavelet transform is employed. Therefore, modelling statistical property of DWT coefficients’ magnitude across scales might request extra investigation with those drawbacks in mind. Further arguments and details about this matter of DWT descriptors will be discussed in the section 2.
Figure 1: Wavelet (solid line) & Wavelet Packet Decomposition (solid and dash lines) for 1-D signal
5.2. Discrete Wavelet Packet Transform

Well-known DWT can be computed efficiently by an orthonormal FIR conjugate quadrature filter banks \( g_0, g_1 \) including analysis low-pass and high-pass filters (denoted \( \phi_S \) and \( \psi_{s,i} \) respectively. The low-pass coefficients \( c_S(x, y) \) are decomposed recursively for a number of levels, and inverse discrete wavelet transform is calculated by an inverse filter bank. The extension from normal wavelet transform (DWT) to wavelet packets transform (DWPT) is straightforward by an additional step at each processing level. Instead of decomposing only low-pass coefficients in the Low-Low sub-band for 2D input signals, the transform performs decomposition of high-pass coefficients in Low-High, High-Low and High-High sub-bands as well. As a result, all coefficients of DWPT can be neatly arranged in a binary tree and addressed as follows.

\[
d_{s,i}(x, y), \ s \in [0, S], \ i \in [0, 4^s - 1]
\]

where \( s \) is a analysis depth in the tree, \( S \) notes the deepest decomposed level, and \( i \) is the node index in this depth. With regards to other representations, wavelet packets have advantages in their adaptability to varying statistical structure. Unlike Fourier Transform with one fixed-size base or normal wavelet transform with a fixed number of bases, we may search the “best” orthonormal bases from dictionary of basis acquired after wavelet package decomposition. This idea is initially proposed by Coifman et al.\cite{28} mainly for signal compression. Therefore, this “best” basis is the best in terms of compressing ratio which often desires sparsest representation. In other words, input signals can be characterized by few large coefficients. Supposed the whole best basis operation is denoted as \( B^2 \) which exhaustively goes through the whole binary tree to look for locations of a set basis with parameters \( (s, i) \) such that the there is a minimum amount of uncertainty measured by Shannon entropy. More details can be found in Coifman’s works \cite{28}, and \( B^2 \) can be summarized mathematically as follows.

\[
B^2(d_{s,i}) : (s, i) = \text{argmin}_{s,i} \left( \sum H(d_{s,i}) \right)
\]

Noteworthy that, sometime “brute-force-attack” every branch of the tree is not possible or feasible due to intensive requirement of computational power. Fortunately, there exist fast algorithms to implement the best basis for given signals. Then, the optimum time-frequency representation can be achieved.
by tilting the time-frequency plan in accordance with best-basis algorithms. Though the representation may be optimally sparsest in time-frequency domain, whether sparseness of features suitably matches performance of human visual attention is still a question to be answered. To rectify the matter, experiments have been carried out and performance comparison between the DWT case and DWPTBB case is reported in the section 8.

5.3. Quaternion Wavelet Transform

Like wavelet packet transform in the previous discussion, Quaternion Wavelet Transform (QWT) is extended and enhanced to eliminate shift-variance problems from real discrete wavelet transform (DWT). Though there are a few different definitions and implementations of Quaternion Wavelet [29], the QWT implementation in this paper is inspired by Chan’s research et al [26]. Some backgrounds about complex wavelet transform (CWT) with its implementation dual-tree complex wavelet transform (DT-CWT) Kingsbury et al [30] need reviewing before the QWT can be explained and discussed. In discrete wavelet transform, 2-D DWT can be considered as concatenation of two consecutive 1-D DWTs. Though the same process does not exactly happen in 2-D DT-CWT or QWT straightforwardly. A similar concept is used for easily explaining how QWT can be achieved. It means 1-D DT-CWT will be elaborated first; then, we will discuss about 2-D QWT signals and how it may handle processing along different orientations and sub-bands.

Real DWT have well-known drawbacks in terms of shift-invariance and phases to encode coefficient locations. Kingsbury et al [31] reckons problems and proposes an dual-tree CWT as a specific solution. Rationale of the dual-tree approach is usage of complex numbers for wavelet coefficients which directly tackles one of two DWT’s dragging problems. Complex extension of wavelet transform makes phase extraction from wavelet coefficients possible since complex wavelet transforms have both real and imaginary values unlike real DWT with only one real value for each coefficient. Real and imaginary components of the dual-tree CWT are generated by two sets of wavelet and scaling functions $\psi_h, \psi_g$ and $\phi_h, \phi_g$. Moreover, filter-banks $h_0, h_1$ and $g_0, g_1$ have to be independent and orthogonal as shown in the figure 2.

The notations $\phi_h(x)$ and $\psi_h(x)$ are denoted for scaling and wavelet functions corresponding to filter-banks $h_0, h_1$. In addition, $c_{h_s}$ and $d_{h_s}$ with $s \leq S$ denotes first set of DTCWT coefficients. Similar notations are used for the second set of scaling and wavelet functions $\phi_g(x)$ and $\psi_g(x)$ with filter banks $g_0, g_1$ and according coefficients $c_{g_s}$ and $d_{g_s}$ with $s \leq S$. Wavelet functions
Figure 2: Illustration of DTCWT
ψ_h(x) and ψ_g(x) forms two binary trees in figure 2 and at leaves of each tree is the real and imaginary parts of a complex analytic wavelets.

\[ \psi^{(c)}(x) = \psi_h(x) + j\psi_g(x) \]

Moreover, the imaginary wavelet \( \psi_g(x) \) is 1-D Hilbert Transform of the real-wavelet \( \psi_h(x) \):

\[ \psi_g(x) = \mathcal{HT}(\psi_h(x)) \]

Any complex wavelet coefficient are formed by wavelet coefficients of two other real wavelet transforms; therefore, this combination generates a \( 2 \times 2 \) redundant tight frame. This redundancy in complex wavelet frames prevents non-oscillating magnitudes of coefficients around singularity points as well makes the transform near-shift invariant. Furthermore, there is no energy (or little energy in practice) in the negative region of frequency because of a relationship between two wavelet functions \( \psi_h \) and \( \psi_g \).

\[
\Psi_g(\omega) = \begin{cases} 
-j\Psi_h(\omega) & \omega > 0 \\
 j\Psi_h(\omega) & \omega < 0 
\end{cases}
\]

then

\[
\Psi^c(\omega) = \mathcal{F}(\psi^c(x)) \\
= \mathcal{F}(\psi_h(x) + j\psi_g(x)) \\
= \Psi_h(\omega) + j\Psi_g(\omega)
\]

and

\[
\Psi^c(\omega) = \begin{cases} 
2\Psi_h(\omega) & \omega > 0 \\
0 & \omega < 0 
\end{cases}
\]

Thus, the Fourier transform of complex wavelet transform \( \Psi^c(\omega) \) has no energy in the negative frequency region. It makes DTCWT an analytic wavelet transform with analytic output signals. Due to this analyticity, the dual-tree wavelet transform has implicitly managed to include all information in the half positive plan of the frequency domain.

It is quite straightforward for 2-D DWT expansion from 1-D DWT, discussion in the previous subsection. However, it is unfortunately not easy for similar expansion from 1-D DTCWT to 2-D DTCWT transforms because Hilbert Transform (HT) and analytic signals need an theoretical extension.
for 2-D signals. Furthermore, there exists not only one but several definitions which define different zero-out regions (negative frequency domain in 1D DT-CWT), signal-power regions (positive frequency domain in 2D DT-CWT). In this paper, we only focus on Bulow definition \cite{Bulow} of analytic quaternion signals which combines both partial and total Hilbert transform (HT). Partial HTs are done along either $x$ or $y$ directions only; meanwhile, total HT is carried out on both directions simultaneously. They are defined as following formula.

\[
\begin{align*}
 f_{H_1} (x) &= f(x) \circ \frac{\delta(y)}{\pi x} \\
 f_{H_2} (x) &= f(x) \circ \frac{\delta(y)}{\pi x} \\
 f_{H} (x) &= f(x) \circ \frac{1}{\pi^2 xy}
\end{align*}
\]

The $f_{H_1}, f_{H_2}, f_H (x)$ are partially Hilbert transformed along $x$ and $y$ axis consequently, and $f_H (x)$ is total HT; while $\circ$ denotes 2-D convolution. Each 2-D CWT basis is a complex analytic function, computationally equivalent to a product of two 1-D complex wavelet functions either along only one or both axis. Similar to expansion of discrete real wavelet, the diagonal sub-band wavelet is defined as $f(x) = \psi_h(x) \psi_h(y)$. Other total and partial HT are products of coefficients from different sets of wavelet functions deployed in the 1-D CWT implementation.

\[
(f_{H_1}, f_{H_2}, f_H) = (\psi_g(x)\psi_h(y), \psi_h(x)\psi_g(y), \psi_h(x)\psi_g(y))
\]

To unify all different Hilbert Transform in a meaningful and compact representation, we can utilize quaternion algebra and treat $f(x)$ as a real part and $(f_{H_1}, f_{H_2}, f_H)$ as three imaginary components \cite{Chan}.

\[
f_A^q(x) = f(x) + j_1 f_{H_1}(x) + j_2 f_{H_2}(x) + j_3 f_H(x)
\]

More details about theory behind QWT and its special characteristics such as its singular cases, three phases, and zero-out regions can be found in Chan et al.’s and Bulow’s publications \cite{Chan, Bulow}. Resting on form of the above quaternion wavelet transformation, we can organize four quadrant components of 2-D wavelet $(f, f_{H_1}, f_{H_2}, f_H)$ as a quaternion. Lets take an example of diagonal signals with following quadrant components.

\[
(f, f_{H_1}, f_{H_2}, f_H) = (\psi_h(x)\psi_h(y), \psi_g(x)\psi_h(y), \psi_h(x)\psi_g(y), \psi_g(y)\psi_g(y))
\]
We can have a diagonal quaternion wavelet functions for the diagonal sub-band mathematically defined as follows.

\[ \psi^D(x, y) = \psi_h(x)\psi_h(y) + j_1 \psi_g(x)\psi_h(y) + j_2 \psi_h(x)\psi_g(y) + j_3 \psi_g(x)\psi_g(y) \]

To compute the QWT coefficients, we can use proposals of a separable 2-D implementation of dual-tree filter-banks previously illustrated in the figure. At each filtering stage, both two-sets of wavelet filters \( h \) and \( g \) are independently applied to each dimension \( x \) and \( y \) of a 2-D image. For example, the filter-bank \( h \) is applied along both axis; then, it yields the scaling coefficients \( c_{hh} \) and three diagonal, vertical and horizontal wavelet coefficients \( d_{hh}^D, d_{hh}^V, \) and \( d_{hh}^H \) respectively as shown in the figure. Dual-tree implementation of two separated filter-banks for 1-D signal can be considered as four independent filter banks for 2-D signals according to all possible combinations of filter for one dimension \( (hh, hg, gh, gg) \). With these combinations of filters and corresponding wavelet functions \( \psi(x)\phi(y), \phi(x)\psi(y) \) and \( \psi(x)\psi(y) \) are generated four components of quaternion wavelet transform for horizontal, vertical, and diagonal sub-bands. Four different wavelet coefficients from these filter banks are arranged by quaternion algebra to obtain QWT coefficients. For example, a coefficient from diagonal wavelet sub-band of QWT
can be written in terms of responses from independent filter-banks as follows.

\[ d_s^D = d_{hh}^D + j_1 d_{gh}^D + j_2 d_{hg}^D + j_3 d_{gg}^D \]

So far, we have taken a diagonal sub-band as example for showing how QWT can be computed. The construction and properties for other two sub-bands are similar to what have been done for diagonal sub-bands. Except that the axis combinations results in a horizontal sub-band \( \psi(x) \phi(y) \) or for a vertical sub-band \( \psi(x) \psi(y) \) instead of a diagonal sub-band \( \psi(x) \psi(y) \). In summary, QWT at each stage sports three quaternion sets corresponding to three sub-bands; each quaternion contains four wavelet functions. Therefore, there are 12 functions in total which can be easily seen as matrix of functions as follows.

\[
\begin{bmatrix}
  d_h^H & d_h^V & d_h^D \\
  d_{gh}^H & d_{gh}^V & d_{gh}^D \\
  d_{hg}^H & d_{hg}^V & d_{hg}^D \\
  d_{gg}^H & d_{gg}^V & d_{gg}^D \\
\end{bmatrix}
= q
\begin{bmatrix}
  \psi_h(x) \phi_h(y) & \phi_h(x) \psi_h(y) & \psi_h(x) \psi_h(y) \\
  \psi_g(x) \phi_h(y) & \phi_g(x) \psi_h(y) & \psi_g(x) \psi_h(y) \\
  \psi_h(x) \phi_g(y) & \phi_h(x) \psi_g(y) & \psi_h(x) \psi_g(y) \\
  \psi_g(y) \phi_g(y) & \phi_g(y) \psi_g(y) & \psi_g(y) \psi_g(y) \\
\end{bmatrix}
\]

Columns of the above matrix correspond to quaternion wavelet functions of the horizontal sub-band \( d_h^H \), the vertical sub-band \( d_h^V \), and diagonal sub-band \( d_h^D \) from left-to-right respectively. The three according wavelet coefficients are \( d_s^H \), \( d_s^V \) and \( d_s^D \) and the \( q \) operator means formation of quaternion number by coefficients along each column. Though quaternion wavelet coefficients possess rich phase information, our research currently focuses on magnitudes of each wavelet sub-bands. Therefore, magnitudes of horizontal, vertical sub-bands can be computed according to quaternion magnitude formula as follows.

\[
\| d_s^H \| = \sqrt{(d_{hh}^H)^2 + (d_{gh}^H)^2 + (d_{hg}^H)^2 + (d_{gg}^H)^2}
\]

\[
\| d_s^V \| = \sqrt{(d_{hh}^V)^2 + (d_{gh}^V)^2 + (d_{hg}^V)^2 + (d_{gg}^V)^2}
\]

\[
\| d_s^D \| = \sqrt{(d_{hh}^D)^2 + (d_{gh}^D)^2 + (d_{hg}^D)^2 + (d_{gg}^D)^2}
\]

While the final approximated version of input signals, which are not decomposed further by the transform, have its magnitude computed by quaternion
algebra.

\[ \|c_S\| = \sqrt{(c_{hh})^2 + (c_{gh})^2 + (c_{hg})^2 + (c_{gg})^2} \]

\[ = \sqrt{(\phi_h(x)\phi_h(y))^2 + (\phi_g(x)\phi_h(y))^2 + (\phi_h(x)\phi_g(y))^2 + (\phi_g(y)\phi_g(y))^2} \]

5.4. Quaternion Wavelet Packet Transform

To construct a packet form of QWT, each and every sub-band \(c_S, d_H, d_V, d_D\) should be repeatedly decomposed by low-pass \((h_0, g_0)\) and high-pass filters \((h_1, g_1)\). Bayram et al \[32\] has investigated into formation of wavelet packets for DT-CWT, an equivalent form of QWT. In order to get an analytic quaternion wavelet packet, the filter banks need to be chosen in a specific way such that the Hilbert transform relationship is preserved. In Bayram’s works \[32\], the analytic wavelet transformed can be achieved if whatever filter-bank is used to decompose the first filter-bank of QWT should also be used for the second (dual) filter-banks. Another important point about the extension to wavelet packet QWPT is the choice of the extension filters \(f_i(x)\).

It has been found that the only necessary constrain to preserve the Hilbert transform property is forcing the usage of the same filter-pairs \(f_0(x), f_1(x)\) in both filter-banks of QWT or DT-CWT. Therefore, any CQF pair of filter-banks with short support, frequency selectivity or possessing a number of vanish moments can be candidates for the extension filter. Noted that, the above criteria such as CQF pair of filters have been employed for extending a regular DWT. Like other derivatives of DT-CWT or QWT, the quaternion wavelet packet transform (QWPT) are approximately shift-invariant, which means the energy in each sub-band is approximately preserved if the input signals are shifted by a number of samples. Noteworthy that, there are other methods beside QWPT with shift-invariant property in wavelet packet decomposition. For example, by performing an exhaustive search over all shifted wavelet packet bases to find the “best basis” according to a certain cost function \[33\], the orthonormal wavelet packet transform becomes shift-invariant in a sense that energy in each sub-band is invariant to transition of input signals. This (approximately) shift-invariance property becomes very useful and important in the search for a suitable energy descriptors. This shift-invariance property guarantees that DT-CWT, DT-CWPT or QWT, QWPT would have energy descriptors robust to certain amount of affine transformation in input signals.

Interactions of filtering both low and high components at each stage of DT-CWT introduces a complete structures of all possible sub-bands that
can be generated by the filter-bank pair. Each tree forms a unique frequency
profile of input signals. Among those countless numbers of possibilities, there
exist a frequency decomposition being more sparse and compact than the
others. It is called the best-basis in terms of representing the input signals
with fewest wavelet coefficients. A fast algorithm for indicating such best
basis has been reported in extension from DWT to DWPT [28]. In addition,
it is previously mentioned in the section 5.2, the same strategy can be adopted
for searching best-basis in QWT. In brief, the approach is looking for a path
in a binary of decomposition to minimize a Shannon entropy cost function;
more details can be found in the work of Coifman et al [28]. After “best
basis” searching for QWPT decomposition, we can identify magnitudes and
energy of coefficients at a specific location by a simple quaternion algebra.

\[ \|q(a, b, c, d)\| = \sqrt{a^2 + b^2 + c^2 + d^2}. \]

6. Wavelet Coefficients Correlation

The previous section 4 have discussed the potential of using energy den-
sity distribution of localized time-scale element or wavelet elements instead
of pixel-value probability distribution. Only general 1-D signal is considered
and these elements are assumed to be independent or at least linearly in-
dependent (uncorrelated); however, this assumption only works for random
variables as input signals. Practically, except total noise, any meaningful sig-
als often has specific structures persistent across multiple time-scale element
in 1-D case. For 2-D signals like natural images, an additional orientation
needs considering; in other words, their wavelet coefficients are highly sta-
tistically related across scales, orientation, spaces. This phenomenon is sys-
tematically studied and confirmed in Azimifar et al research [34]. The author
has conducted an empirical study of joint wavelet statistics for texture and
natural images to investigate correlation relationship between neighbouring
coefficients. Examination of these dependencies helps propose appropriate
models for such a transform-domain algorithm. Though Azimir’s work [34]
only covers linear dependencies and just a squint on non-linear relations, its
proposals are evaluated on a collection of 5000 real images. Therefore, we
believe her conclusion in that study is generally true at least for natural im-
ages, the main researching objects. In brief, there exists a few elementary
correlation relationships as follows.

- The spatially-localized and sparse correlation structure has a clear per-
sistence across scales.
Every coefficient exhibits correlations extending across multiple scales, with spatially near neighbors both within and across orientations.

- A subband coefficients at the same spatial locations but from different orientations are not linearly correlated.

- Within-subband, inter-subband, and inter-scale correlations are highly oriented and persistent across local neighbors of its parent.

The below figure 4 clearly illustrates all mentioned correlations, their preferences to locality as must be expected. This locality increases toward finer scales, which supports persistency property of wavelet coefficients. A single coefficient correlates with its parents as well as neighbors across orientations and scales. Among several mentioned statistical dependencies, the most vi-

Figure 4: Illustration of wavelet coefficients inter-band and intra-band correlation [34]

...tal findings for our work are uncorrelated siblings coefficients across orientation and strong correlated coefficients across scales since it theoretically allows uncertainty and mutual information estimation of a 2D time-scale element or a wavelet sub-band energy descriptor. To elaborate this point, lets consider two adjacent scales $s_1, s_2$ and their corresponding coefficients $w_i f(x, y, s_i)$ of horizontal, vertical, diagonal orientation $i = v, h, d$ for 2-D signals $f(x, y)$. Due to non-correlation of sibling coefficients across orientation, it is possible to consider three wavelet coefficients as a multivariate variable $W_s = (w_h, w_d, w_v)$ with uncertainty estimation by energy density distribution across three orientations $H(W_s)$. For two adjacent sub-band
1, s2, there are two multivariate variables Ws1 and Ws2 corresponding two
entropy values H(Ws1) and H(Ws2), and the mutual information between
two variables due to inter-scale dependencies between correspondent wavelet
coefficients are computed as follows.

\[ I(W_{s1}, W_{s2}) = H(W_{s1}, W_{s2}) - H(W_{s1}) - H(W_{s2}) \] (6)

From that basic observation about inter-scale and intra-scale wavelet coef-
ficients of natural images is developed the core idea of our proposal. More
details about sub-band energy descriptors and how to measure their uncer-
tainty and mutual information will be clearly explained in following sections
6.1, 6.2

6.1. Interscale Subband Energy Descriptor

Interesting relationship between basis-project methods and scale saliency
are repeatedly discussed in several publications [7], [13]. Kadir [7] actually dis-
cusses about behaviours of non-saliency and saliency regions in spectral and
wavelet domain. A simple, flat, non-salient regions or images is sufficiently
described by a single sub-band; meanwhile, complicated data and structure
regions require more sub-bands descriptors. This directly introduces basis-
projected sub-bands as potential alternative descriptors. Like pixel-value
descriptors, real wavelet sub-bands must be treated as discrete variables due
to its theoretical restriction, data-analysis uncertainties, \( \sigma_t \sigma_\omega \geq \frac{1}{2} \). In other
words, it is impossible for continuous wavelet sub-bands distribution at any
specific location. Following available mathematical definition of PSS for dis-
crete pixel descriptors, we sketch rough mathematical models of WSS with
discrete sub-band descriptors, \{ e \in E, E = \{ e_1, e_2, \ldots e_m \} \}, in the equations
7, 8, 9, 10 whereof e, E are a element and set of sub-band descriptor consec-
tutively.

\[ Y_D(s_p, \bar{x}) \triangleq H_D(s_p, \bar{x})W_D(s_p, \bar{x}) \] (7)
\[ H_D(s, \bar{x}) \triangleq -\sum_{d \in D} p_{b,s,\bar{x}} \log_2 P(d, s, \bar{x}) \] (8)
\[ W_D(s, \bar{x}) \triangleq \frac{s^2}{2s - 1} \sum_{d \in D} |p_{b,s,\bar{x}} - p_{b,s-1,\bar{x}}| \] (9)
\[ s_p^- \triangleq \{ s : H_D(s - 1, \bar{x}) < H_D(s, \bar{x}) > H_D(s + 1, \bar{x}) \} \] (10)

However, a general concept of sub-band descriptor is not useful in actual com-
putation; therefore, an appropriate numerical attribute of sub-bands need
proposing instead. Let's consider 2-D discrete real-wavelet transform with three sub-bands vertical (\(v\)), horizontal (\(h\)) and diagonal (\(d\)) sub-bands at each particular dyadic scale \(s\) represented by three set of wavelet coefficients \(w_i\) accordingly in the equation \(11\). Equation \(12\) uses those coefficients to compute sub-band energy densities as descriptors \((e)\) for Wavelet-domain Scale Saliency (WSS).

\[
\begin{align*}
    w_i f(x, y, s_j) |_{i=\{v, h, d\}, s_j=\{s_1, s_2, \ldots, s_n\}} &= f(x, y) \ast \psi_{s,i}(x, y, s_j) \quad (11) \\
    P\{w_i f(x, y, s_j)\} |_{i=\{v, h, d\}, s_j=\{s_1, s_2, \ldots, s_n\}} &= |w_i f(x, y, s_j)|^2 \quad (12)
\end{align*}
\]

In the standard real discrete wavelet transform (DWT), there are fixed three analysed sub-bands for each dyadic sampling step. Supposedly the maximum level of wavelet decomposition is \(n\), the number of dyadic scales is \(n\) with 3 sub-bands for each scale. With 4 or 5 as the usual number of decomposition levels, totally around 12 or 15 sub-bands descriptors are analysed for an image. This number of descriptors is significantly less than 255 pixel-value descriptors of PSS for any grey-scale image.

Besides wavelet transforms, different other types of basis projection techniques could also be utilized; for example, best basis wavelet packet analysis (DWPTBB). The full wavelet packet transform breaks signals into sub-bands with the same bandwidth at the maximum dyadic scale. It would not fit into the scale saliency concept which requires descriptors at different scales. Fortunately, the "balanced" full wavelet packet tree usually over-describes image properties, and the description can be optimized by Best Basis \((B^2)\) finding operation. The optimized wavelet packet tree often has projected basis across dyadic scales since some small image details are best described with a basis at finer resolution while other big details prefer another basis with coarser resolution. The DWPTBB coefficients are utilized for sub-band energy density, the proposed image descriptors, calculation in the equations \(13\) \(14\).

\[
\begin{align*}
    w_i f(x, y, s_j) |_{(i, j)=B^2(w_i f(x, y, s_j))} &= f(x, y) \ast \psi_{s,i}(x, y, s_j) \quad (13) \\
    P\{w_i f(x, y, s_j)\} |_{(i, j)=B^2(w_i f(x, y, s_j))} &= |w_i f(x, y, s_j)|^2 \quad (14)
\end{align*}
\]

Comparing mathematical statements \(11\) \(12\) and \(13\) \(14\) for sub-bands descriptors of Discrete Wavelet Transform (DWT) and Discrete Wavelet Packet Transform Best Basis (DWPTBB) consecutively, we can see their fundamental differences. While DWT provides determinant basis-projection methods with pre-computed basis and fixed structure of sub-bands, DWPTBB adapts
itself into each data set. Then, its number and structure of sub-bands are
specified by Best Basis ($B^2$) finding operator [28]. It requires more opera-
tions; however, more faithful and adaptive descriptors can be achieved.

Both DWTBB and DWT are popular wavelet-transforms; however, they
both depend on shift-variant real discrete wavelet transforms. It means that
projection of coefficients not only depends on data but also its relative location
on the scene.

$$w_1f(x, y, s_j) \neq w_1f(x + \Delta(x), y + \Delta(y), s_j), \exists x, y, s_j, w_1, \Delta(x), \Delta(y)$$

(15)

$$P\{w_1f(x, y, s_j)\} \neq P\{w_1f(x + \Delta(x), y + \Delta(y), s_j)\}, \exists x, y, s_j, w_1, \Delta(x), \Delta(y)$$

(16)

As the fourth criteria for good information measurement of Starck et al.[17]
states that entropy must work in the same way regardless of descriptors’ lo-
cations. Both DWT and DWTBB projected descriptors do not satisfy that
condition since usages of these descriptors might lead to different information
estimation for identical data at two different locations. The shift-variance
of real-wavelet transform can be avoided by complex wavelet transform de-
design; for instances, recently developed dual-tree complex wavelet transform
(DTCWT) [35], Quaternion wavelet transform (QWT) [26], or dual-tree com-
plex wavelet packet transform with best-basis (DTCWTBB) [32]. General
formula of complex coefficients and their corresponding sub-band energy den-
sity are summarized in the equations [17][18]

$$w_1f(x, y, s_j)|_{i=\{v,h,d\}\vee B^2(w_1f)} = f(x, y) * (\psi_{g,s,i}(x, y, s_j) + j\psi_{h,s,i}(x, y, s_j))$$

(17)

$$P\{w_1f(x, y, s_j)\}|_{s_j=\{s_1,\ldots,s_n\}\vee B^2(w_1f)} = \|w_1f(x, y, s_j)\|_2^2$$

(18)

Dual-tree approaches use two different wavelet filter-banks, \{\psi_g, \psi_h\}, and
they are designed to form analytical complex filter banks, \{\psi_g(x, y) + j\psi_h(x, y), \psi_h(x, y) \approx
H (\psi_g(x, y))\}. The magnitudes of projected-complex coefficients are proven
to be shift-invariant; therefore, its derived energy density of the sub-bands
is as well shift-invariant. Probably, the quaternion version of wavelet trans-
form (QWT) and quaternion wavelet packet transform best basis (QWTBB)
with shift-invariant property would provide better descriptors than their real
counterparts according to five criteria of Starck [17].
6.2. Intra-scale Subband Energy Descriptor

As previously mentioned in the section 6, there is strong correlation or statistic linear dependence between wavelet coefficients in natural images. The first correlation, the inter-scale dependencies, has been discussed in the section 4 and modeled as sub-band descriptors in the previous section 6.1. Moreover, the relation have been widely and effectively employed in various tree-structured coding techniques such as SPIHT [36]. Besides inter-scale relationship, many authors [25] have pointed out another strong correlation of intra-band coefficients existing across many different types of natural scenes. Minh Do and M Vetterli [25] successfully modelled coefficients of a wavelet sub-band with a simple explicit mathematical form, Generalized Gaussian Distribution (GGD). While statistical distribution of wavelet coefficients gets a lot of interest, several researches have proposed different mathematical models for analysing this statistical characteristic. However, few models of wavelet coefficients marginal density at a particular sub-band works better than GGD in terms of accuracy, approximation and simplicity. After such the distribution is widely observed in experimental data with natural images, theoretical analysis on the plausibility of modelling by the GGD distribution is defined as follows.

\[
p(x; \alpha, \beta) = \frac{\beta}{2\alpha \Gamma(1/\beta)} e^{(-|x|/\alpha)^\beta}
\]

where \( \Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt \), \( z > 0 \) is the Gamma distribution. Here \( \alpha \) dictates the scale parameter or variance of the distribution, and \( \beta \) controls shapes. For example, GGD with \( \beta = 1 \) is Gaussian distribution; it becomes Laplacian distribution with \( \beta = 2 \).

7. Information Measurement

In the previous section 6.1, four different wavelet transforms generate corresponding wavelet sub-band energy density descriptors. From those energy density, energy probability distribution function (PDF) at each scale \( s_j \) can be computed as follows.

\[
p_{\text{interband}}(x, y, s_j) = \frac{p\{P[w_i f(x, y, s_j)]\}}{\sum_i P[w_i f(x, y, s_j)]}
\]
The above formula computes the probability of energy density at one location \((x, y)\) across different sub-bands, \(i = \{v, h, d\}\) for (DWT) or (QWT) or \(i = B^2(w_i)\), for (DWPTBB) and (QWPTBB) from the smallest scale, 1, to currently considered scales, \(m\). The first level uses the smallest sampling window size of wavelet atoms; therefore, it generates analysed coefficients with finest details. Then, the sampling window sizes are doubled after each level; they generate coarser analysed details. It is quite similar to PSS sampling operations except that scales are doubled rather than increased by a unit. Like PDF of PSS descriptor, WSS descriptors PDF are distributed with increasing scales of \(j\), from level 1 (smallest wavelet atom) to level \(m\) (currently biggest wavelet atom). From the equation 19, it is straightforward to compute feature-space entropy \(H_{\text{Observer}}(x, y, s_m)\) as follows whereof \(p\{P[w_i f(x, y, s_j)]\}\) is shorted as \(p_{\text{interband}}(x, y, s_j)\).

\[
- \sum_{\{i=\{v, h, d\}\land i=B^2(w_i)\}, \{j< m\}} p_{\text{interband}}(x, y, s_j) \log p_{\text{interband}}(x, y, s_j) \tag{19}
\]

Both entropy of PSS’s descriptors and the above entropy formula for the proposed descriptor only summarizes statistical property in local spatial regions since both considering window sizes in PSS and scale levels of wavelet decomposition are finite. Then, it lacks involvement of energy distribution in the whole image and it is confirmed that such distribution is vital for natural image and texture modeling [25]. As presented in the sub-section 6.2 is the Generalized Gaussian Distribution of coefficients magnitudes from a wavelet intra-band.

\[
p_{\text{intraband}}(x, y, s_j, \alpha, \beta) = \frac{\beta}{2\alpha \Gamma(1/\beta)} e^{-\sqrt{x^2+y^2/\alpha}} \tag{20}
\]

In order to combine both global and local characteristics into a single value, we propose cross-entropy \(H_{\text{Searcher}}(x, y, s_m)\) between inter-band and intra-band distribution as an alternative formulation of the equation 19.

\[
- \sum_{\{i=\{v, h, d\}\land i=B^2(w_i)\}, \{j< m\}} p_{\text{interband}}(x, y, s_j) \log p_{\text{intraband}}(x, y, s_j) \tag{21}
\]

To distinguish between two modes of entropy computation, we name the local entropy by the equation 19 as "observer" mode, and the cross-entropy involving both local and global statistics as "searcher" mode. In later
formula, when general entropy symbol $H$ without specific subscripts appears in any formulas, it means both modes are eligible for those equations. Those names also help to distinguish different parameters and simulation modes presented in the experimental sections 8.2.

The equation 19 computes feature-space entropy of sub-band energy descriptors for WSS as the equation 8 does for PSS. Half of scale saliency measure, feature-space entropy, has been figured out for sub-band energy density descriptors. The other half of the problem rests in computational details of inter-scale saliency; in other words, how the equation 9 should be interpreted with the proposed descriptors. In equation 8, the inter-scale saliency is measured as total variation in probability distribution of descriptors at two consecutive scales in which pixel-value descriptors ($d$) appear in both distributions, it complicates the problem. However, the situation is different for wavelet sub-band energy density descriptors since each sub-band in the current level is unique for this level only. It does not appear in other levels of analysis. This wonderful property simplifies out task in building sub-band probability distribution for different levels but makes the equation 3 inappropriate for sub-band features. Since it is unjustifiable to find total variation of two $PDF$ on two different set of descriptors, an alternative interpretation of inter-scale saliency need developing. Lets consider

$$P(M) = \{p_{i,j}(x,y,s_{j})|\forall i,j \leq m\}$$

$PDF$ of all sub-bands up to the current level, $m$. When a new analysed sub-band, $D = \{p_{i,j}(x,y,s_{j})|j = m + 1\}$, is generated, this sub-band descriptor will modify the current $PDF$ into $P(M|D)$. The distance between the prior model and the modified model can be measured by Kullback-Leibler divergence as follows.

$$K(P(M|D), P(M)) = \int_{M} P(M|D) \log \frac{P(M|D)}{P(M)}$$

(22)

Noteworthy, it is similar to Itti’s Bayesian Surprise Saliency (BSS) metric [14], and the surprise model can be extended for multiple sub-bands descriptors or evidences in BSS. The equation 22 becomes mutual information between the current model and a set of new evidences. In other words, the expectation of surprise for adding new sub-bands into the current model is the mutual information between new sub-bands and the current model, shown in the equation 23

$$MI(D, M) = \int_{D} K(P(M|D), P(M)) = \int_{D,M} P(D, M) \log \frac{P(D, M)}{P(M)P(D)}$$

(23)
Therefore, mutual information is chosen as inter-scale saliency for successive dyadic scales since it actually implies averaged ”bayesian surprise” \cite{14} saliency of sub-bands across scales. Furthermore, mutual information as inter-scale saliency measurement well emphasizes the structural coherence of data across scales. If there are useful structures such as edges or joints and they are consistent across consecutive scales, they will increase mutual information between two consecutive scales. Otherwise noises have no mutual information across scales as its self-information is zero, \( I(N,N) = 0 \). It is remarkable that mutual information satisfies the fifth criterion of the good information estimation by Starck et al.\cite{17}. The only remaining step is identifying how the mutual should be calculated in discrete cases. Following formula shows relation between mutual information and entropy.

\[
MI(D, M) = H(D) + H(M) - H(D, M) \tag{24}
\]

\[
H(M) = - \sum_{\{i=(v,h,d)\cup i=\mathbb{B}^2(w_i), \{j\leq m\}\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{25}
\]

\[
H(D) = - \sum_{\{i=(v,h,d)\cup i=\mathbb{B}^2(w_i), \{j=m\}\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{26}
\]

\[
H(D, M) = - \sum_{\{i=(v,h,d)\cup i=\mathbb{B}^2(w_i), \{j\leq m+1\}\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{27}
\]

The mutual information can be directly calculated as difference between separated \((H(D) + H(M))\) and joint \((H(D, M))\) entropy estimation of the current energy descriptors (the current model) and the next-level sub-bands, the equation \cite{24}. While the entropy elements \(H(D), H(M)\) and \(H(D, M)\) can be easily estimated by simple mathematical equations \cite{25, 26, 27}. The joint entropy \(H(D, M)\) can be reused as \(H(M)\) for the next level inter-scale saliency estimation because of the sub-band descriptors uniqueness. The scale saliency principles on wavelet-domain sub-band energy descriptors are summarized in the equation \cite{28} as product of maximum feature-space saliency and inter-scale saliency, or product of mutual information between consecutive levels and maximum sub-band entropy.

\[
H(M(x, y, s_p)) = - \sum_{i=(v,h,d)\cup i=\mathbb{B}^2(w_i), j\leq m} p_i(x, y, s_p) \log p_i(x, y, s_p)
\]

\[
MI(D(x, y, s_p), M(x, y, s_p - 1)) = H(D) + H(M) - H(D, M)
\]

\[
s_p \triangleq \{ s : H(M(s - 1, x, y)) < H(M(s, x, y)) \land H(M(s, x, y)) > H(M(s + 1, x, y)) \}
\]

\[
Y(M(x, y, s_p)) = H(M(x, y, s_p)) \ast MI(D(x, y, s), M(x, y, s_p - 1)) \tag{28}
\]
The characteristic scale \( s_p \) is chosen to maximize information of the model \( H(M(s, x, y)) \). Let’s imagine the case prior scale contains only noise meanwhile later scales actually contain useful structures of images. With bias of Shannon entropy toward noise, the characteristic scale fails to enclose any useful structure. To overcome this drawback, we propose alternative approach, DIS to differentiate from the original strategy WSS, in which \( s_p \) is selected so as to maximize inter-scale saliency or average ”Bayesian surprise”. DIS principles can be summarized as follows.

\[
s^* = \{ s : MI(D_{s-1}, M_{s-2}) < MI(D_s, M_{s-1}) \land MI(D_s, M_{s-1}) > MI(D_{s+1}, M_s) \}
\]

Experiments with DIS and WSS are carried out and simulations results are detailed in the next section in order to confirm effectiveness of the proposed strategy.

8. Discussion & Results

The previous sections 6.1 and 7 have analysed theoretical advantages of WSS and its derivative DIS. In addition, the subsection 6.1 present four descriptors based on different wavelet transforms: DWT, QWT, DWPTBB, and QWPTBB. Accordingly, we have several derivatives for the proposed method according to specific choices of scale section mechanisms and sub-band descriptor. To evaluate them against other saliency approaches, they are compared with PSS [7], and the de-facto ITT model [3]. The purpose of comparisons are not for claiming the best saliency method or racing toward the highest possible evaluating measurement; it just proves the rationale of the assumption that feature and structural complexity would be a good clues for human attention. The best evaluation measurement reported does not necessarily mean the best saliency maps since it much depends on choices of databases, elimination of experimental bias, performance of human test subjects, etc. Moreover, a standardized evaluation process in saliency map evaluation is far from being reached since several researchers choose different database and measurement methods or even create their own. In our research, we focus on the effectiveness of information measurement in visual attention; then, the most common processes and databases would be chosen to confirm generalization of the assumption.

In line of searches for informative clues, Bruce and Tsotsos [15] database is certainly among the popularly used stimuli. However, only Bruce’s database
is certainly not enough due to limits in numbers and contents of stimuli. Then, Kootstra’s database [16] are chosen for extra testing samples and ground-truths. Two database with over 200 samples with ground-truths provided by more than 50 human subjects would help to confirm the generalization of our proposed framework to a certain extent. Similarly, only common evaluation approaches are deployed in our studies, and they can be categorized into either quantitative or qualitative methods. Quantitative relations between different saliency methods and human visual performance are shown by appropriate statistical methods (AUC, NSS) with eye-tracking data as ground-truths. Meanwhile, the qualitative results, visual comparisons of different saliency maps, gives a glimpse about performance for each individual sample. It also specifies imaging contexts where saliency methods give reasonable solution as well as situations where saliency maps are unreasonable to human perception.

8.1. Databases of image stimuli

The ground-truth and data for basic evaluations of visual saliency performance is got from eye-tracking experiments. Specially in Neil Bruce database, 120 different color images are observed in random orders while there are 4 seconds gap between the previous and the next stimuli. To ensure consistency and accuracy of the database, subjects are asked to seat 0.75 m in front of a 21 inch CRT monitor. Especially, human subjects have no further instructions for any actions or clues for what images appear next. Furthermore, image contents are varied from indoor to outdoor environments. Sometimes, there are clear interesting objects in the scene; while some scenes are really general without any particular interests in any subjects. A non-head mount eye tracking apparatus extracts locations of eye-fixation while human test subjects look at sample images. Other setting-up parameters are intended for a general-scene based stimuli which are typically found in urban environments. Moreover, the same parameters are used for collecting data from 20 different subject over 120 testing samples. The following figure 5 shows first eight images from the Neil Bruce’s database. Despite its popularity, images from Neil Bruce’s database has narrow semantic content since it contains only urban scenes and mainly indoor environments. Besides that, the number of samples are relatively small. Due to that, an additional database should be included in simulations such that there more testing images of natural objects like animals, flowers, in natural environments. Kootstra’s database [16] have just satisfied these requirements with
additional ground-truths for further experiments. Kootstra’s ground-truths
data are also collected from eye-tracking experiments although the exper-
imental process is slightly different from what have been done to collect
Bruce’s database. In the psychological experiment, with head-mount eye-
tracking devices, thirty-one students (15 men, 16 women) ranging from 17
to 32 of age took part in the experiments, and they are all naive about aims
of experiments. Each human subject observes a total of 99 photographic
image in five different categories while their eye movements are recorded si-
multaneously with the head-mounted device. There are nineteen images in
natural symmetry category; each of which contains symmetrical natural ob-
jects. Beside such symmetrical scenes, other non-symmetrical photographic
scenes are included into the image sets such as: 12 images of animals in nat-
ural setting, 12 images of street environments, 12 images of street scenes, 16
images of building and 40 images of natural environments. Figure 6 gives an
eample of 5 categories of images in the Kootstra’s database. Noted that,
each image is presented to viewers with a resolution of 1024x768 pixels on
an 18” CRT monitor at a distance of 70 cm from the participant.

8.2. Quantitative Comparisons of Saliency Methods
The quantitative performance includes Receiver Operating Characteris-
tics (ROC) curves with Area Under ROC Curve (AUC), and Normalized
Scanpath Saliency (NSS) as numerical results. To ensure fair comparisons
between methods, open-source evaluation codes for AUC and NSS [37] are
employed. Noteworthy, saliency maps are standardized around median in-
stead mean of distributions. Quantitative evaluation of visual saliency map
on natural images with eye-tracking data ground has been initially studied by Tatler and recently summarized by Borji et al. More information about mechanisms behind ROC and AUC can be found. In this section, we only focus on usages of these quantitative methods to compare, evaluate and prove rationale of our approach. As the main purpose of this evaluation is confirming effectiveness of informative clues in human’s visual attention, our approach is not optimally tuned to reach the maximum AUC or NSS.

All four descriptors mentioned in the section have been simulated with image samples from both Neil Bruce’s and Kootstra’s datasets. Noted that scale selection mechanisms have strong influences in formation of saliency maps; therefore, two separated simulations are carried out to investigate that effect as well. Figure 7 and 8 summarizes simulations results of proposed methods with corresponding WSS and DIS respectively in Neil Bruce’s image dataset.

According to the figure 7 and the table 2, performances of four WSS derivatives follow decreasing orders: DWT, QWT, DWPTBB and QW-PTBB; however, all are better than PSS performance and comparable to ITT method. Especially, a computational time is deducted by approximately 7 times; noteworthy, the PSS is implemented in C++ with MATLAB interface and WSSs are totally written in MATLAB. For Niel Bruce database, the best
Figure 7: ROC Curve - WSS

| MTH    | AUC   | NSS   | TIME(s) |
|--------|-------|-------|---------|
| ITT    | 0.6944| 0.27714| 1.096s  |
| PSS    | 0.5856| -0.39175| 7.1092s |
| DWT    | 0.67823| 0.33358| 1.2401s |
| QWT    | 0.66279| 0.30002| 1.9231s |
| DWPTBB | 0.6417| 0.26079| 2.6187s |
| QWPTBB | 0.63529| 0.23714| 5.2836s |

Table 2: Quantitative Result
basis approaches $DWPTBB$ and $QWPTBB$ does produce poorer results in both accuracy test, AUC and NSS as well as efficiency test, TIME.

![ROC Curves](image)

Figure 8: ROC Curve - DIS

| MTH     | AUC       | NSS       | TIME(s) |
|---------|-----------|-----------|---------|
| ITT     | 0.6944    | 0.27714   | 1.096s  |
| PSS     | 0.5856    | -0.39175  | 7.1092s |
| DWT     | 0.7028    | 0.3178    | 1.2689s |
| QWT     | 0.6922    | 0.3024    | 1.9527s |
| DWPTBB  | 0.6299    | 0.2546    | 2.4218s |
| QWPTBB  | 0.6394    | 0.2351    | 5.4835s |

Table 3: Quantitative Result

Quantitative performances of four DIS methods are shown in the figure 8 and the table 3. Mixed results are spotted. Performances of DWT and QWT descriptors with DSS approach are a little bit increased in terms of AUC if compared to the case of WSS. However, ”Best-basis” descriptors ($DWPTBB,QWPTBB$) perform a little bit better if WSS are employed instead of DIS. Meanwhile, there is almost no difference between WSSs and DSSs in term of both NSS and TIME regardless descriptors.
Above is shown simulation results from Neil Bruce image data-set with eye-tracking locations. Despite of its recently popular database in evaluating saliency maps, the data-set has limitations analysed in the subsection 8.1. Another sets of images should be brought in to enhance diversity of testing samples. Kootstra’s database with more image categories and all eye-tracking data ground truth is a perfect candidate. Additional simulation results would confirm and generalize rationale of our proposed information-based saliency methods. Similar to the table 2, figure 7, the table 4 and figure 9 demonstrate how well the proposed methods with four descriptors and WSS scale selection mechanism perform against other saliency methods like ITT and PSS.

![ROC Curves](image)

**Figure 9: ROC Curve - WSS**

| MTH       | AUC     | NSS   | TIME(s) |
|-----------|---------|-------|---------|
| ITT       | 0.7819  | 0.5144 | 2.3874  |
| PSS       | 0.5852  | -0.3532 | 17.0663 |
| DWT       | 0.7150  | 0.4849 | .4414   |
| QWT       | 0.7301  | 0.5070 | 2.9313  |
| DWPTBB    | 0.7242  | 0.4631 | 4.9577  |
| QWPTBB    | 0.7612  | 0.3922 | 4.7743  |

**Table 4: Quantitative Result**
Among four descriptors, the best result in terms of AUC is from QW-PTBB descriptor, and the second best is QWT; while both DWT and DW-PTBB have nearly equal AUC values. Comparing with ITT and PSS, QW-PTBB’s performance in AUC measurement is nearly equal to that of ITT and much larger than PSS. The result strengthens our hypothesis about usefulness of informative clues in saliency map construction. Moreover, it suggest that sub-band wavelet descriptors would be better pixel-based descriptors for scale-saliency computation. In terms of NSS, QWT has slightly out-performed the other descriptors, and its value nearly approaches NSS result of ITT and obviously surpasses PSS’s result.

The graph and the table shows numeric evaluation of the proposed methods with WSS scale section mechanism on Kootstra’s database. Besides WSS scale selection, we have another method called DIS; therefore, we should compare how DIS performs on the same database with suggested sub-band descriptors. Therefore, similar quantitative assessments are also done for wavelet scale saliency with DIS scale selection and simulation results are shown in figure and table.

![ROC Curves](image)

**Figure 10: ROC Curve - DIS**

With this specific simulation parameters, the method still performs quite well against other methods like ITT and PSS in both terms of AUC and
Table 5: Quantitative Result

| MTH       | AUC   | NSS   | TIME(s) |
|-----------|-------|-------|---------|
| ITT       | 0.7819| 0.5144| 2.3874  |
| PSS       | 0.5852| -0.3532| 17.0663 |
| DWT       | 0.7058| 0.4847| 0.4169  |
| DTCWT     | 0.7173| 0.5027| 2.8972  |
| DWPTBB    | 0.7173| 0.4556| 4.8156  |
| DTCWPTBB  | 0.7381| 0.3468| 4.7152  |

NSS. However, both AUC and NSS of DIS are slightly worse than those of DIS methods. Noteworthy, ITT analyses three channels color, intensity and orientation simultaneously while we just utilize a intensity channel. Only one channel is chosen since we try to isolate the performance of wavelet scale saliency from other external effects such richness of input features. Regardless of DIS or WSS, the method has very competitive results in numeric terms and it is not due to comprehensiveness of input features.

8.3. Qualitative Comparison of Saliency Methods

In this section, we show a few examples of visual saliency maps from two mentioned database of Bruce and Kootstra. From each of the databases, only four test images are chosen to be displayed due to limited space though saliency maps are generated for every single image in either of the databases. The samples are intentionally chosen to show variety of contexts and scenes as well as they cover cases of successfully highlighting interested objects and cases of failing to emphasize salient regions. Along with the proposed methods, saliency-maps of ITT and PSS methods are also included so as to give visual comparisons to our proposed saliency methods. Directly below are displayed four samples from Neil Bruce’s database.

There are four samples, shown in figures 11,12,13,14, for qualitatively analysing. Generally, PSS identifies a large portion of images as salient regions (white regions), it explains why its average AUC and NSS in the table 2 are the lowest, and ITT model gives reasonable saliency maps for three over four samples. Four samples of saliency maps are deliberately chosen to show that different ranking of WSS, DIS derivatives, and their dependence on mother wavelet morphological shapes. Sometimes, their performances are quite similar, the figure 11, however, QWT-WSS performs better than DWT-WSS in many samples; for example, figure 7.
Figure 11: Saliency Map 1
Figure 12: Saliency Map 2
Figure 13: Saliency Map 3
Figure 14: Saliency Map 4
QWPTBB-WSS also have their own advantages, especially in the case text-
tured background - figure 13. Finally, sometimes none of proposed methods
do give reasonable saliency maps, figure 14. It usually happens if images are
flooded with complex textures.

While Neil Bruce’s data-set capture daily scenes in the urban and subur-
ban areas, it lacks of scenes from natural landscapes. Therefore, its images
do not represent the whole meaning of “natural images” category. In order
to visually confirm effectiveness of our proposed methods, we include four
”natural” samples with corresponding saliency maps from the Kootstra’s
database in the following figures 15, 16, 17 and 18.

![Figure 15: Saliency Map 1](image)

In accordance with one sample image (in color), we display saliency maps
produced by ITT, PSS, and eight derivatives of the proposed wavelet scale
saliency methods. The figures 15 and 18 represents image of flowers taken in
close distance. Therefore, it shows quite a number of symmetric and small
details. Meanwhile, figures 16 and 17 contains the whole landscape of moun-
Figure 16: Saliency Map 2
Figure 17: Saliency Map 3
Figure 18: Saliency Map 4
tains and plateaus. Those scenes are usually anti-symmetric and much richer in information than flowery photos. In general observations, ITT method does the best job in selecting the salient features. Among derivatives of the proposed features, QWPTBB descriptors show the most competitive and comprehensive visual result, followed by QWT, DWTBB, and DWT based derivatives in descending order of performances. For comparison between WSS and DIS scale-selection mechanisms, there is slight but significant difference between their saliency maps - WSS saliency maps in the second rows and DIS saliency maps in the third rows of the figures [15][16][17] and [18]. In these examples, DIS maps tend to highlight more features than those of WSS; in other words, WSS maps might have better discriminant power than DIS maps. It would explain why AUC and NSS results in the table 4 are slight better than those in the table 5. There are small changes in quantitatively visual results when different parameters are used. However, the proposed methods performs very well against other saliency methods like ITT and PSS.

9. Conclusion

In this paper, we propose the extension of scale saliency from pixel descriptors to sub-band energy density descriptors generated by four DWT, DWPTBB, QWT, and QWPTBB wavelet transforms with two different scale selection mechanisms WSS and DIS. Comparing to pixel-value descriptors (PSS), the proposed descriptors are much more sparse but biased toward morphological shapes of mother wavelet. Moreover, the proposed descriptors are more robust to external influencing factors to generation of saliency maps such as shift-variance and other affine transformation. Furthermore, wavelet packet descriptors with best basis algorithms are also considered since several psychological experiments suggest sparseness factor in human vision system. Along with new descriptors, innovative coherent information framework for wavelet scale saliency is proposed and strong relations with Bayesian Surprise Model [14] are emphasized. Beside solid theoretical development, the experimental results are as well competitive with state-of-the-art ITT model and surpasses the original scale saliency model PSS quantitatively and qualitatively. In future research, theoretical analysis will be extended to include prior information or top-down information, perceptual grouping and other visual attention operations.
References

[1] U. Neisser, Cognitive Psychology, 4319, Appleton-Century-Crofts, 1967.

[2] David Marr, Vision: A Computational Investigation into the human representation and processing of visual information, Inc, New York, NY, 1982.

[3] L. Itti, C. Koch, E. Niebur, Others, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on pattern analysis and machine intelligence 20 (1998) 1254–1259.

[4] D. Walther, Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics, interactions 2006 (2006).

[5] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, Advances in neural information processing systems 19 (2007) 545.

[6] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society, 800, Citeseer, 2007, pp. 1–8.

[7] T. Kadir, M. Brady, Saliency, scale and image description, International Journal of Computer Vision 45 (2001) 83–105.

[8] S. Gilles., Robust Description and Matching of Images, Ph.D. thesis, University of Oxford, 1998.

[9] N. Bruce, J. Tsotsos, Saliency based on information maximization, Advances in neural information processing systems 18 (2006) 155.

[10] D. Gao, V. Mahadevan, N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency, Advances in Neural Information Processing Systems 20 (2007) 1–8.

[11] L. N. Anh Cat, G. Qiu, U. Geoff, A. Li-minn, S. Kah Phooi, Visual Information Based on Fast nonparametric multidimensional entropy estimation, International Conference on Acoustic, Speed and Signal Processing (2012).
[12] G. Qiu, X. Gu, Z. Chen, Q. Chen, C. Wang, An information theoretic model of spatiotemporal visual saliency, in: international conference on multimedia and expo, Citeseer, 2007, pp. 1806–1809.

[13] T. Kadir, D. Boukerroui, M. Brady, An analysis of the scale saliency algorithm, OUEL No: 2264 3 (2003) 1–38.

[14] P. Baldi, L. Itti, Of bits and wows: A Bayesian theory of surprise with applications to attention., Neural networks : the official journal of the International Neural Network Society 23 (2010) 649–66.

[15] N. D. B. Bruce, J. K. Tsotsos, Saliency, attention, and visual search: An information theoretic approach, Journal of Vision 9 (2009) 1–24.

[16] G. Kootstra, B. de Boer, L. R. B. Schomaker, Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry., Cognitive computation 3 (2011) 223–240.

[17] J. Starck, F. Murtagh, Multiscale entropy filtering, Signal Processing 76 (1999).

[18] P. Suau, F. Escolano, A new feasible approach to multi-dimensional scale saliency, in: Advanced Concepts for Intelligent Vision Systems, Springer, 2009, pp. 77–88.

[19] D. Stowell, M. D. Plumbley, Fast multidimensional entropy estimation by k-d partitioning, Signal Processing 16 (2009) 537–540.

[20] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum Neurobiol 4 (1985) 219–227.

[21] F. Urban, B. Follet, C. Chamaret, O. Meur, T. Baccino, Medium Spatial Frequencies, a Strong Predictor of Salience, Cognitive Computation (2010).

[22] A. B. Watson, The cortex transform: Rapid computation of simulated neural images, Computer Vision, Graphics, and Image Processing 39 (1987) 311–327.
[23] H. Senane, A. Saadane, D. Barba, Image coding in the context of a psychovisual image representation with vector quantization, in: Proceedings., International Conference on Image Processing, volume 1, IEEE Comput. Soc. Press, 1995, pp. 97–100.

[24] M. Unser, N. Chenouard, D. Van de Ville, Steerable pyramids and tight wavelet frames in L2(R(d))., IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 20 (2011) 2705–21.

[25] M. N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation., IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 14 (2005) 2091–106.

[26] W. L. Chan, H. Choi, R. G. Baraniuk, Coherent multiscale image processing using dual-tree quaternion wavelets., IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 17 (2008) 1069–82.

[27] S. Mallat, A wavelet tour of signal processing, 2, Academic Pr, 1999.

[28] R. R. Coifman, M. V. M. Wickerhauser, Entropy-Based Algorithms for Best Basis Selection, IEEE Transactions on Information Theory 38 (1992) 713–718.

[29] T. Bulow, G. Sommer, Hypercomplex signals—a novel extension of the analytic signal to the multidimensional case, IEEE Transactions on Signal Processing 49 (2001) 2844–2852.

[30] N. Kingsbury, The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters, Proc. 8th IEEE DSP Workshop (1998) 2–5.

[31] N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals, Applied and Computational Harmonic Analysis 10 (2001) 234–253.

[32] I. Bayram, On the Dual-Tree Complex Wavelet Packet and M-Band Transforms, Processing, IEEE Transactions on (2008) 1–27.
[33] I. Cohen, S. Raz, D. Malah, Orthonormal shift-invariant wavelet packet decomposition and representation, Signal Processing 57 (1997) 251–270.

[34] Z. Azimifar, M. Amiri, P. Fieguth, E. Jernigan, Empirical Study of Wavelet Domain Image Joint Statistics and Proposition of an Efficient Correlation Map, Journal of Mathematical Imaging and Vision (2011).

[35] I. Selesnick, R. Baraniuk, N. Kingsbury, The dual-tree complex wavelet transform, IEEE Signal Processing Magazine 22 (2005) 123–151.

[36] R. Buccigrossi, E. Simoncelli, Image compression via joint statistical characterization in the wavelet domain, IEEE Transactions on Image Processing 8 (1999) 1688 –1701.

[37] A. Borji, D. N. Sihite, L. Itti, Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study, IEEE transactions on image processing : a publication of the IEEE Signal Processing Society PP (2012) 1.