Deep Learning Algorithm for Video Saliency Object Detection using 3D DWT with Set Partition Integer Hierarchical Tree List

Suresh Babu D[1] and Cyril Prasanna Raj[2]

[1]Research Scholar Dept of ECE, UVCE, Bangalore University, India
[2]Professor, Dept of ECE, CIT, Bangalore. India

Abstract: In this work detection of salient objects in image and video sequences with higher accuracy, faster processing speed and reduced computation complexity is designed using deep learning algorithm. 3 Dimensional (3D) Discrete Wavelet Transform is combined with Set Partition Integer Hierarchical Tree List (SPIHTL) encoding methods for identifying self-similarity coefficients of the salient object to be detected. Deep learning algorithm with seven layers is proposed in this invention that processes the quantized wavelet sub bands obtained after 2-level 3D inverse DWT reconstruction process detects the optimum number of features representing salient objects in the input data. An apparatus for evaluation of proposed method for salient object detection is designed in this invention that reconstructs the image from the features. Evaluation metrics are identified for measuring the performance of the proposed methods in this invention and are compared with traditional methods. The measure of PSNR and MSE for 120 different data sets of various dimensions and orientations demonstrates an improvement of 12%-18% in PSNR measurement as compared with existing methods.

Keywords: Salient objects, deep learning, 3D DWT, SPIHT, 2D FFNN

1. Introduction

Human visual system is very unique system than can recognize the objects of interest in an image or video sequence and then start identifying the neighbouring objects to corroborate finer details of the object of interest extracting high level of information from the scene. Scientists have been studying this cognitive property of the human visual system for computer vision application development such as scene understanding. Research studies on detecting important regions and objects in an image data effectively has been carried out by cognitive scientists over last few years in particular for computer vision applications. Detection and recognition of objects, compression of image and video data, imaging process such as photo collage [1], cropping, thumb nailing, quality assessment of video and image data, image retrieval based on content [2], browsing of net based on image data [4], tracking of objects [5], human robot interaction and object discovery [6] are few of the applications of salient object detection process. Salient object detection is generally classified into bottom-up or top-down approaches [3]. Salient objects are those that are of important in a given scene, objects that have already been recorded that appears in a given scene [7], objects in cluttered scene, objects that are very interesting, surprise objects, aesthetic, attributes and scene context. Most of the salient models need to consider neurobiological aspect as the objects that will be considered will have motion associated in video sequences [8]. Detection salient objects with motion require understanding of neuroscience. Studies on video saliency detection algorithms have been reported by Liu et al [9] for multimedia applications. Traditional video saliency detection methods compute salient objects from each frame in the spatial domain and salient features from temporal domain.

With resurgence of neural networks and deep learning with advantages of independency on features and bias information salient object detection is carried out using Convolutional Neural Networks (CNN) [10]. In CNN model there are thousands of neurons with tunable parameters that can identify salient regions using receptive fields. In deep learning methods based on CNNs salient object detection is categorized into two main methods: classical CNNs and Fully Convolutional Networks (FCN). In CNN models classification networks process the segment information obtained from multi-layered perceptrons that extract features in several scales. In

Received: January 29th, 2021. Accepted: September 25th, 2021
DOI: 10.15676/ijeei.2021.13.3.7

FCNs instead of considering sub images or patches in the input image as in CNNs every pixel is considered even near the boundaries of salient objects. With fully CNNs demonstrating superiority in terms of performance over traditional methods of salient object detection process there is lot of research studies in this direction [10]. If there are transparent objects in the scene or if there is low contrast between video frames, detection capability of FCN based algorithm is limited [11]. In the studies reported in [12,13] filtering of images prior to salient object detection will improve reliability and accuracy as there will be noise during data acquisition if the images or video sequences are acquired in real time. Challenges in salient object detection in video sequences are the constraints that exit in intensity changes or contract variations between video frames and the motion of objects between frames [14]. Guanqun Ding, Yuming Fang [15] in their work have presented 3D convolutional networks for extracting spatio-temporal features from video sequences and 3D de-convolutional networks for fusing spatio-temporal features extracted for salient object detection. To enhance the edge features wavelet transform is used as pre-processing layer for salient object detection [16]. Wavelet transform of image data decomposes the input image into multiple sub bands of low frequency components that capture the DC components or intensities of input image, and all other sub bands capture the high frequency components or the edge information in the input images [17]. The high frequency sub bands localize the edge information along the vertical, lateral and diagonal axis providing flexibility in processing and edge enhancement process [18]. Combining wavelets with neural networks auto encoders have been designed comprising of three layers for classification of objects [19]. CNNs process the featured enhanced objects for salient object detection; however with both forward and inverse wavelet transform requiring additional processing time due to complexity in arithmetic operation, it is required to carry out salient object detection in wavelet domain. In this work, fully connected deep learning model that process data in the wavelet domain is proposed for salient object detection in both image and video sequences.

2. Related work

Studies have been reported in literature on use of wavelet sub bands for training deep neural networks for classification process and object detection. Combining Support Vector Machine (SVM) and k-Nearest Neighbour (KNN) with wavelet features is used for handwritten recognition [20]. Wavelets are used with CNNs as pre-processing layer for classification of images, detection of textures and for improving face resolution [21]. Wavelet sub bands are combined or fused prior to classification [22] or wavelet sub bands are used to compute feature vectors for classification [23] or significant features are extracted from wavelet feature for classification [24]. Liu et al. have presented algorithms for image restoration that combines CNN with multi-level wavelet sub bands. De Silva et al [16], in their work have presented mechanisms to enhance edge information in the wavelet domain and then perform classification process using CNNs. The high frequency or detail wavelet sub bands are only considered for edge enhancement that is carried out using gradient algorithms and modulus maxima methods. Inverse wavelet transform is carried out to reconstruct the image after edge enhancement without the approximation coefficients. Salient object detection methods based on wavelet transforms and deep learning have been demonstrated to achieve higher accuracy. With wavelet sub bands being used for enhancement of edge features, training of CNNs have improved efficiency in terms of object detection and classification. For video data the temporal domain features correlated with spatial domain features have demonstrated accuracy in salient object detection process using deep learning methods over that of traditional methods. The input images or video sequences are downscaled to standard size of 32 x 32 prior to processing by deep machine algorithms; this dimensionality reduction has significant impact on objects in the image and causes resolution issues. With wavelet transform sub bands being used for enhancement of edges and features and further performing inverse wavelet transform to obtain reconstructed image has improved feature enhancement prior to deep leaning. However, the process of performing DWT and inverse DWT leads to time consumption and computation

complexity. The DWT sub bands capture information in the low pass and high pass sub bands that are half the size of original image. Processing these sub bands that represent intensity components and high frequency components or edge information in the low pass and high pass sub bands respectively reduces computation time by 50%. Enhancement of edge features in the wavelet sub bands have been achieved by using gradient operation and maximum modulus methods as mentioned previously. These methods are also time consuming and also have limitations in terms of eliminating features that are closely co-correlated with objects and edges of the objects. In order to overcome these limitations novel techniques have been proposed in this work that are based on improvising existing methods of salient object detection using wavelets and deep learning algorithms. The proposed work is compared [25 – 27] in relevant aspects and the results are concluded.

3. Proposed Method

In the proposed method of salient object detection that is developed for detecting salient objects in both image and video sequences 3D DWT algorithm is used to capture the intensity and high frequency features from both spatial and temporal into multiple sub bands. The proposed algorithm is presented in Figure 1. The input video frame of size N x N x 64 is decomposed into m-levels using 3D DWT (m is set to 1 or 2 or 3). If m is set to 1, there will be eight sub bands that are ordered in a sequence. In order to retain the wavelet coefficients that are very significant and represent the salient objects, a modified method is proposed in this work based on Set Partition Hierarchical Integer Trees (SPHIT) algorithm. The modified SPIHT algorithm retains only the significant wavelet coefficients that are highly correlated to the intensity levels of the objects in the low frequency band. The encoded wavelet sub bands are reordered from 3D to 2D and are processed by the proposed 2D FFNN structure (with multiple layers and neurons designed to process data without reordering of input data from 2D to 1D). The 2D FFNN structure is designed to detect the objects of interest or salient objects and classify the objects. The output of 2D FFNN achieves dimensionality reduction of detected features that can be used for classification of objects. Further the features can be processed to detect salient objects.

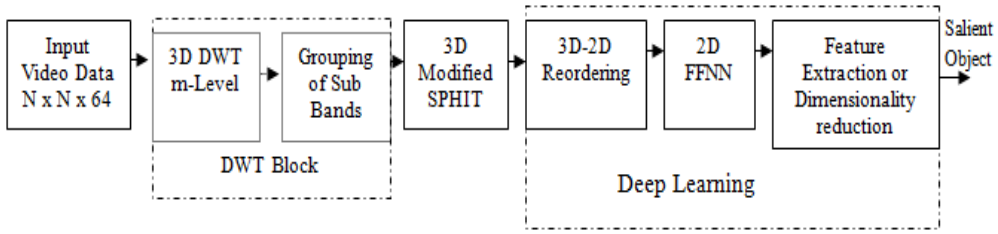

Figure 1. Proposed block diagram for salient object detection in video sequences

The proposed algorithm has two major blocks that are 3D processing and deep learning modules. The modified SPIHT algorithm proposed in this work interfaces the two processing blocks for salient object detection without loss of information.

A. DWT algorithm

DWT is the process of decomposing the input image into multiple sub bands that capture the low frequency and high frequency components in the input image in half the resolution. In this work Daubechies 4 and 9/7 filters are used for computing 3D DWT sub bands. In 3D DWT, there are three modules that process the input image along the x, y and z direction. Figure 2 presents the building block of 3D DWT (17). In each block the low pass and high pass filter is used to process data either in the x-direction (rows) or y-direction (columns) or the z-direction (temporal). The input data or video frame of size N x N x 64 is decomposed into eight sub bands each of size N/2 x N/2 x 32 sub bands.

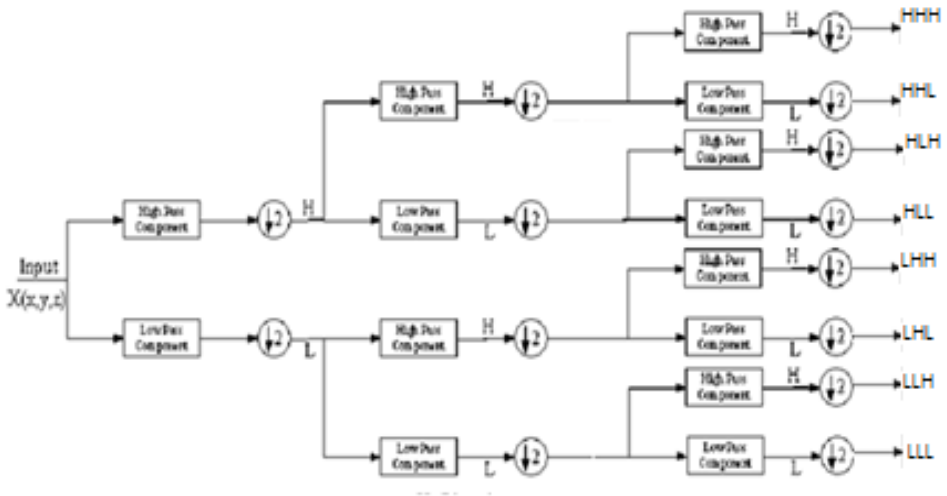

Figure 2. DWT processing of video frames

The first stage process the input data (each frame) along the rows to decompose each frame into 2 sub bands of L and H each of size N x N/2. The second stage processes these two sub bands along the columns by two pairs of filters to further decompose each of these sub bands into two sub bands each of size N/2 x N/2. The two stage decomposes each frame into four sub bands each of size N/2 x N/2. With 2D DWT operating on each of the 64 frames, generates 256 sub bands each of size N/2 x N/2. The 3D DWT decomposition generates eight sub bands represented as LLL, LLH, LHL, LHH, HLL, HLH, HHL and HHH. The LLL sub band captures the intensity component in each of the frame as well as the intensity component in the temporal direction. The LLH component captures the intensity level in each frame and the changes in these intensities in the temporal direction. The LLL sub band is further decomposed into second level DWT sub bands with 3D DWT and third level DWT sub bands as shown in Figure 3. There will be 10 sub bands after 3 level decomposition of N x N x 8 input data.

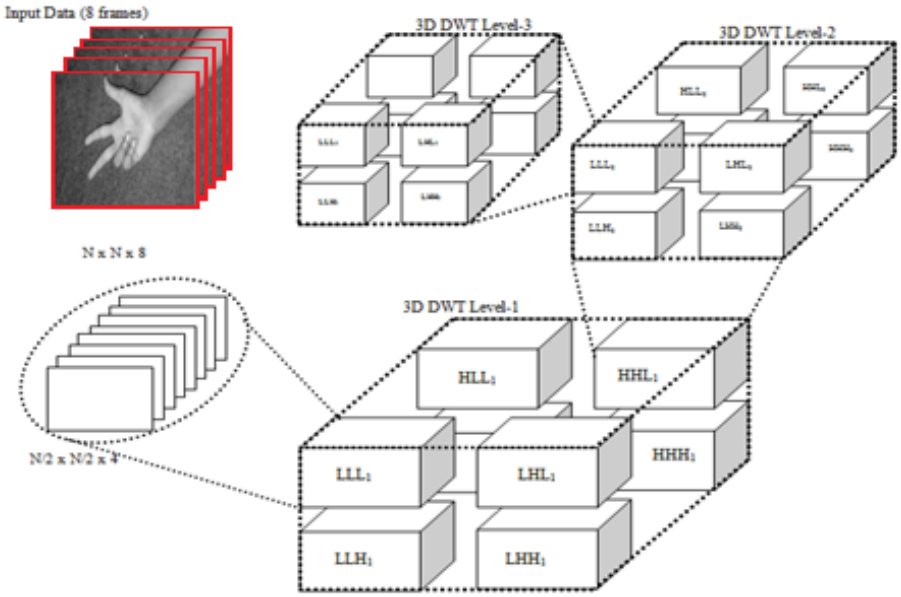

Figure 3. Three-level 3D DWT decomposition

Considering all these sub bands will add to computation complexity, similarly there are edge features that are captured in higher level sub bands of level 2 and level 1 those are also very important for improving salient object detection accuracy. One of the important aspects in 3D DWT decomposition is the redundancy in the multi-level sub bands. In this proposed work, all these sub bands generated by the 3-level 3D decomposition are processed for salient object detection from the input video sequence, but only the significant wavelet coefficients that are contributing to salient objects are considered. To consider only the significant wavelet coefficients, modified SPIHT algorithm is considered in this work.

B. SPIHTL Algorithm

The SPIHT algorithm stores the wavelet coefficients in three ordered list such as List of Insignificant Sets (LIS), List of Insignificant Pixels (LIP) and List of Significant Pixels (LSP). The value of each pixel coordinate is stored in the LIP and LSP and the LIS stores the approximation and detail coefficient. The four steps in SPIHT are initialization, sorting, refinement and quantization step update. After multiple iterations the LSP contains the coordinates of the pixels that are evaluated during the refinement pass. In order to identify the significant pixels that are required for salient object detection the SPIHT encoding algorithm is modified only to generate the three ordered list and the encoding of data 1's and 0's is not considered. The wavelet coefficients in the LIP list are set to zero intensity as it is observed that these pixels are not very significant towards information present the in the wavelet pyramid. As the SPIHT algorithm is modified only to identify the location of significant pixels in the ordered tree and is not used for encoding process, the SPIHT algorithm is faster and also retains the pixel objects that are significant and contributing to the information of salient object in the image. The modified SPIHT algorithm is labeled as SPIHT List (SPIHTL) algorithm and is presented in Figure 4 for computing the most significant pixels that contribute towards computing salient object detection process. The SPIHTL algorithm computes the significance of set of coordinates with the Eq. (1) to find the relationship between magnitude comparisons and message bits.

$$S_n(T) = \begin{cases} 1, & \max_{(i,j)\in T}\{|c_{i,j}|\} \geq 2^n \\ 0, & otherwise \end{cases} \qquad Eq. \tag{1}$$

The notations that are used for SPIHTL algorithm representation is as follows:

- *O(i,j)*: set of coordinates of all offspring of node *(i,j)*;
- *D(i,j)*: set of coordinates of all descendants of the node *(i,j)*;
- *H*: set of coordinates of all spatial orientation tree roots (nodes in the highest pyramid level);
- *L(i,j)*: $= D(i,j) - O(i,j)$.

The set partition rules that are used for the modified algorithm as are follows:

1. The initial partition is formed with the sets $\{(i, j)\}$ and $D(i, j)$ for all $(i, j) \in H$
2. If $D(i, j)$ is significant, then it is partitioned into $L(i, j)$ plus the four single-element sets with $(k, l) \in O(i, j)$.
3. If $L(i, j)$ is significant, then it is partitioned into the four sets $D(k, l)$, with $(k, l) \in O(i, j)$.

(I) Initialization: output n = $\lfloor log_2(max_{(i,j)}\{|c_{i,j}|\}) \rfloor$;
set the LSP as an empty list, and add the coordinates $(i, j) \in H$ to the LIP and only those with descendants also to the LIS, as type A entries

(S) Sorting Pass:

S.a) for each entry (i, j) in the LIP do:

S.a.1) output $S_n(i, j)$;

S.a.2) if $S_n(i, j) = 1$ then move (i, j) to LSP

S.b) for each entry *(i, j)* in the LS do:

S.b.1) if the entry of type A then

- Output $S_n(D(i, j))$:
- If $S_n(D(i, j)) = 1$ then
  - For each $(k, l) \in O(i, j)$ do:
    - Output $S_n(k, l)$;
    - If $S_n(k, l) = 0$ then add *(k, l)* to the end of the LIP;
  - If $L(i, j) \neq 0$ then move *(i, j)* to the end of the LIS, as an entry of type B, and go to step S.b.2; otherwise remove entry *(i, j)* from the LIS

S.b.2) if the entry is of type B then

- Output $S_n(L(i, j))$;
- If $S_n(L(i, j)) = 1$ then
  - Add each $(k, l) \in O(i, j)$ to the end of the LIS as an entry of type A:
  - Remove *(i, j)* from the LIS.

Figure 4. SPIHTL Algorithm

For the input data consisting of 4 x 4 elements that is obtained after two-level 2D wavelet decomposition, the SPIHT encoding algorithm encodes the data into three ordered list as shown in Figure 5. After three iterations the LSP list contains significant pixels and the LIP list contains insignificant pixels and there are no insignificant set in the LIS list. Table 1 presents the details of SPIHTL algorithm for an example of matrix of size 4 x 4. In SPIHTL algorithm is quantized by removing the LIP elements to zero, however quantizing all elements will lead to loss of information in the reconstructed image. To overcome these limitations adaptive SPIHTL algorithm is proposed in this work. The encoded data obtained from SPIHTL encoder is decoded and inverse transformation is carried out and the distortions checked considering PSNR parameter with reconstructed image sequence.

LSP

(0, 0) → 26
(0, 2) → 13
(0, 3) → 10
(0, 1) → 6
(1, 0) → -7
(1, 1) → 7
(1, 2) → 6
(1, 3) → 4
(2, 0) → 4
(2, 1) → -4
(2, 2) → 4

LIP

(3, 0) → 2
(3, 1) → -1
(2, 3) → -3
(3, 2) → -2
(3, 3) → 0

LIS

Empty

Figure 5. SPIHTL ordered list

Table 1. Example of SPIHTL algorithm for 4 x 4 matrix

Input Data					SPIHTL Encoded					SPIHTL Decoded				
	0	1	2	3		0	1	2	3		0	1	2	3
0	26	6	13	10	0	26	6	13	10	0	26	6	14	10
1	-7	7	6	4	1	-7	7	6	4	1	-6	6	6	6
2	4	-4	4	-3	2	4	-4	4	0	2	6	-6	6	0
3	2	-2	-2	0	3	0	0	0	0	3	0	0	0	0

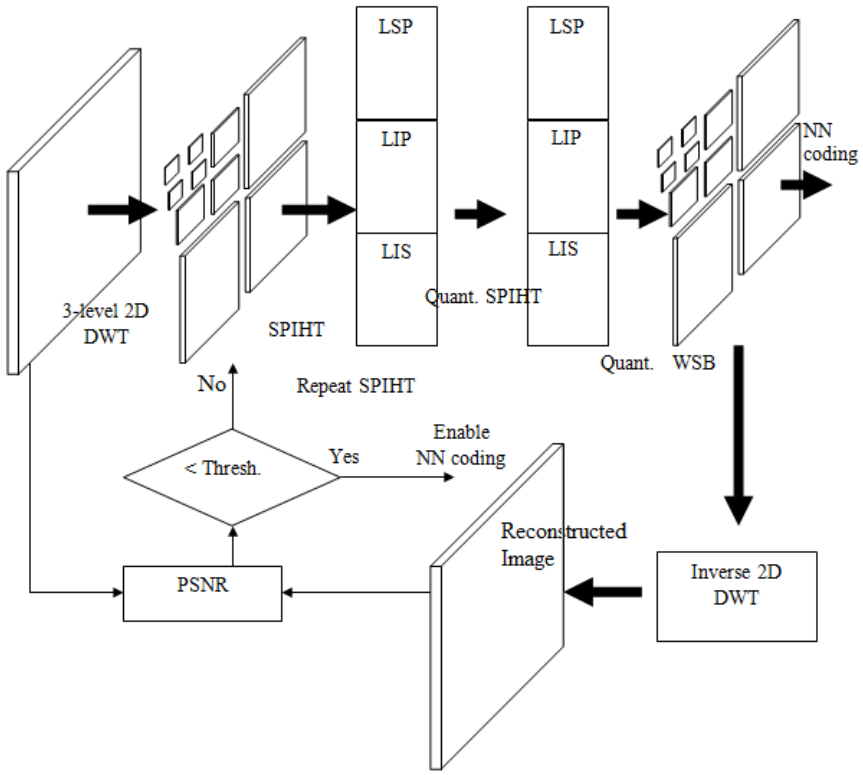

Figure 6. Proposed SPIHT algorithm with novel method of encoding

If the distortion error is very high then the progressive encoding scheme is carried out to the next iteration, this process is continued until the distortion in the reconstructed image is within the set limits. Figure 6 illustrates the proposed algorithm. The number of iterations is set such that every wavelet coefficient is searched for its relevance in the parent-child pair and LIS is empty. The contents of LIP are quantized and made to zero. From the encoded SPIHTL ordered list the wavelet coefficients are rearranged into sub bands. The rearranged sub bands are further processed by the inverse DWT to reconstruct the image. The PSNR is computed considering the original image and reconstructed image, based on the PSNR results and wavelet sub bands are considered for further processing by the NN module for salient object detection. If the PSNR results are less than the desired threshold SPHITL algorithm is revisited to encode the data with threshold level set to 2^{n-1}(16 in this example). If the PSNR are above than the desired results (1.5 time higher than the desired results), SPIHTL is further carried out by setting the threshold level to 2^{n+1} (64 in this example). The SPIHTL algorithm with adaptive logic proposed in this work identifies the pixels in the wavelet domain at higher level and the self-similarity pixels at the all the lower levels that constitute towards salient object detection. The three level decomposed images after quantization process (achieved after performing SPIHTL process) is further processed by the deep learning algorithm for salient object detection and classification.

From the quantized wavelet coefficients two-level inverse 3D DWT is carried out to generate the reconstructed data that generates 8 groups of sub bands each of size N/2 x N/2 x 32 as shown in Figure 7.

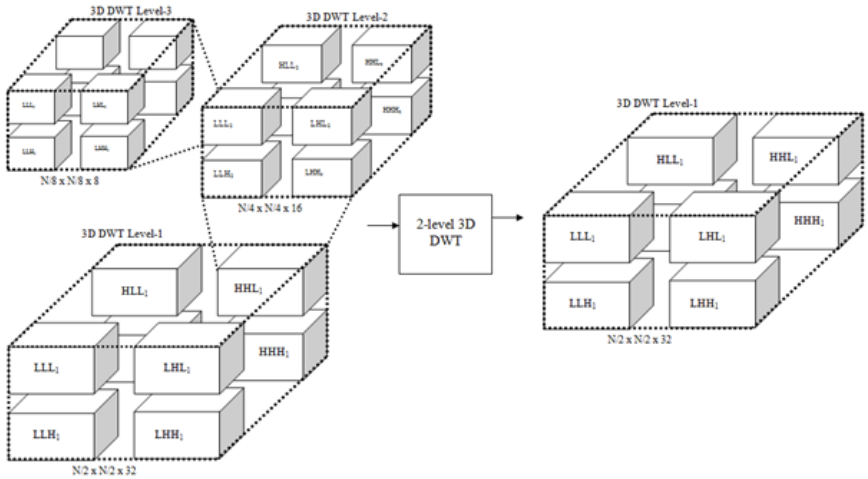

Figure 7. 2-level 3D inverse DWT for reconstruction

After 2-level 3D DWT reconstruction there will be 8 sub bands each of size N/2 x N/2 x 32. The purpose of perfroming 2-level 3D inverse DWT is to ensure that the deep learning algorithm operates on wavelet coefficients to detect salinet objects and classify salinet objects. The advantages of this method is that there are eight sub bands of which one low pass sub band that holds intensity information denoted by (LLL_1) and there are 7 high frequency sub bands denoted by (LLH_1, LHL_1, LHH_1, HLL_1, HLH_1, HHH_1) that are independently processed to detect salient objects. Improvement in accuracy of salient object detection and classification is achieved by considering all the 8 sub bands or only the LLL sub band with additional sub bands selected from any of the seven high frequency sub bands.

C. FFNN for object detection and classification

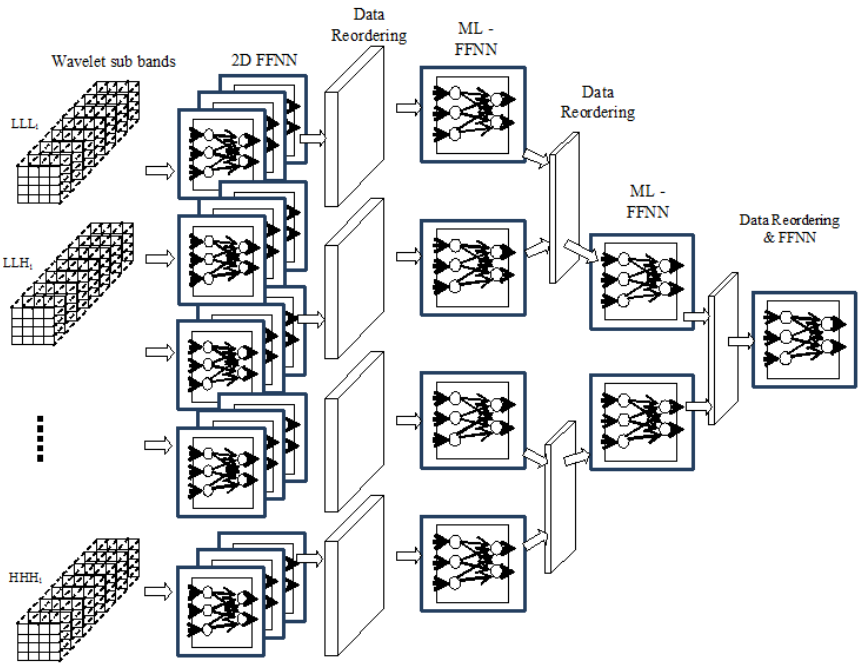

Figure 8. Proposed deep learning structure for salient object detection and classification

The reconstructed image that comprises of 32 frames per sub band with each frame of size N/2 x N/2 and there are eight sub bands that are processed by deep learning structure as presented in Figure 8. The wavelet sub bands are grouped into two sub groups of low pass sub band (LLL_1) and all other sub bands (LLH_1, LHL_1, LHH_1, HLL_1, HLH_1, HHL_1, HHH_1). The LLL1 sub band holds the intensity or DC components of input data and all other sub band holds the directional information along with its motion vector along the temporal direction. For salient object detection the LLL_1 along with any of the other sub bands are considered for processing in the deep learning structure.

The deep learning structure has four multi layered Feed Forward Neural Network (FFNN) structure and three reordering layers. The first stage of FFNN is the 2D FFNN structure, the second, third and fourth FFNN is 1D structure. The three reordered structure are used to rearragned the input elements from 2D to 1D either using zig-zag scanning method or row-column scaning method. Figure 9 presents the internal structure of the proposed 2D FFNN for processing one of the sub images. F^1 and F^2 are the twor consecutive frames that are considered for demonstrating the design of 2D FFNN structure. The sub images of F^1 and F^2 are denoted as $F^1{}_1$ and $F^2{}_1$ respectively. Each of these sub images and its wavelet coefficients are represented as $\{F^1{}_1(0), F^1{}_1(1), F^1{}_1(2), \ldots\ldots F^1{}_1(15)\}$ and $\{F^2{}_1(0), F^2{}_1(1), F^2{}_1(2), \ldots\ldots F^2{}_1(15)\}$.

These 32 coefficients are processed by the 2 x 2 neuron and the intermediatery outputs are denoted as $\{n_1(0), n_1(1), n_1(2), n_1(3)\}$ and the corresponding output of neurons after network activation function processing is denoted as $\{a_1, a_2, a_3, a_4\}$. The relation between the network output a and the network input F is given as in Eqs.(2a) and (2b), similarly all other outputs are represented.

$$n_1(0) = F_1^1(0)W_{1,0}^1 + F_1^1(1)W_{1,1}^1 + F_1^1(2)W_{1,2}^1 + F_1^1(3)W_{1,3}^1 + F_1^1(4)W_{1,4}^1 \ldots\ldots + F_1^1(15)W_{1,15}^1$$

$$F_1^2(0)W_{1,16}^1 + F_1^2(1)W_{1,17}^1 + F_1^2(2)W_{1,18}^1 + F_1^2(3)W_{1,19}^1 + F_1^2(4)W_{1,20}^1 \ldots\ldots + F_1^2(15)W_{1,31}^1$$

$+ b1(0)$ Eq.(2a)

$$a1 = f(n_1(0))$$ Eq.(2b)

The outputs of first layer are reordered and processed by the second layer of FFNN structure.

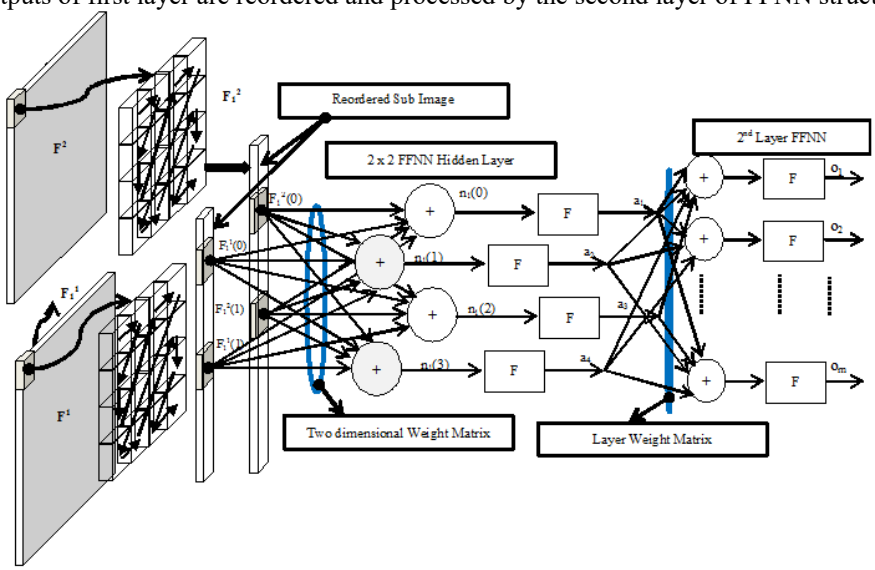

Figure 9. Data processing in the first layer of 2D FFNN structrue

The FFNN structure of second, third and fourth layer are realized using the generic structure as shown in Figure 10, and the outputs of each of the layer is mathematically represented as in Eqs. (3a) and (3b).

$$d_k = \sum_{i=1}^{n} P_i \; W_{k,i} + b_k \,, \; a_k = f(d_k) \qquad \text{Eq.(3a)}$$

$$c_m = \sum_{i=1}^{k} a_k \; W_{m,k} + b_m \,, \; o_m = f(c_m) \qquad \text{Eq.(3b)}$$

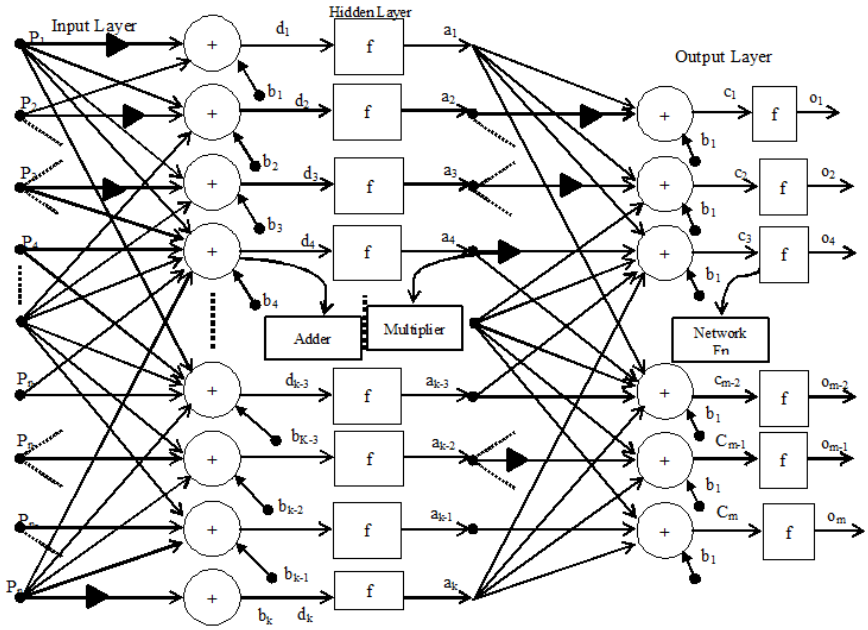

Figure 10. FFNN structure for 2nd, 3rd and 4th layer of deep learning model

The number of neurons in the 2nd, 3rd and 4th layer can be set according to the desing requriements. In this proposed design, the number of neurons in the hidden layer is set to 16 or 8 and the number of neurons in the output layer is set to 4 or 2. The designed FFNN structure reduces the dimensionality of the input data as well as detects the significant features that represent the salient objects.

D. Implementation and Evaluation

The top level block diagram of the deisgned encoder and decoder that is based on 3D wavelet sub band is shown in Figure 11.

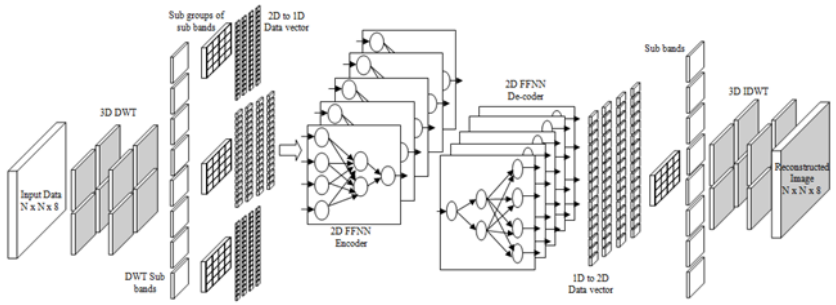

Figure 11. Encoder-decoder modules for salient object detection based on DWT and deep learning

The deep learning network designed is trained considering 30 different objects that are captured using 256 frames during motion. The captured video sequence is grouped into four GOFs each of 64 frames. The images are resized to 512 x 512 resolutions in order to perform DWT and feature selection using modified SPIHT algorithm. By considering 64 frames per GOF, there are 120 GOFs that have 120 different objects captured with motion information. 3-level 3D DWT is carried out and the modified SPIHT algorithm is applied to retain the wavelet coefficients that are very significant towards salient object detection. 2-level 3D inverse DWT is computed and the wavelet sub bands are obtained for training the deep learning network. The inputs are set as the targets at the reconstruction network of deep learning structure. The number of neurons in the encoder module of deep learning network is set to either 4 or 8 or 16. The number of neurons in the hidden layer is adjustable and network activation functions set to both linear and nonlinear. Multiple iterations of training are carried out to evaluate to find the number of neurons for each hidden layer and appropriate network activation functions. The 2D FFNN structure and all the other FFNN structures are modeled in MATLAB environment including the reconstruction network. The training targets are set to achieve minimum gradient or MSE less than 10^{-5}, number of Epochs is set to 600 and number of iterations are set to 1000. The network is trained and the training results are presented in Figure 12 that is used to evaluate the network performances in terms of gradient or MSE achieve and regression number.

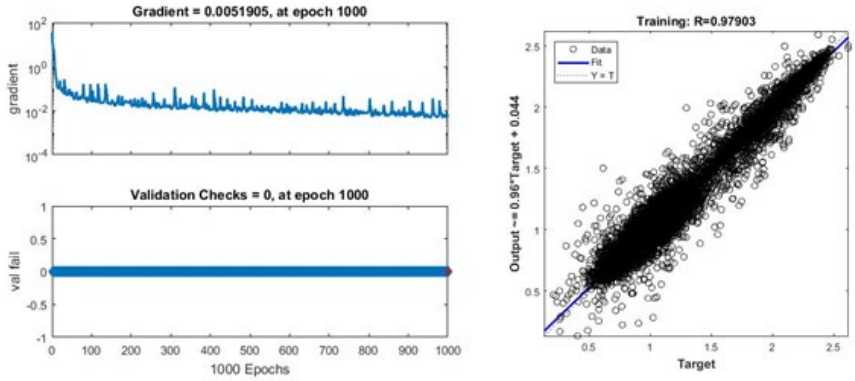

Figure 12. Training performance of deep learning network

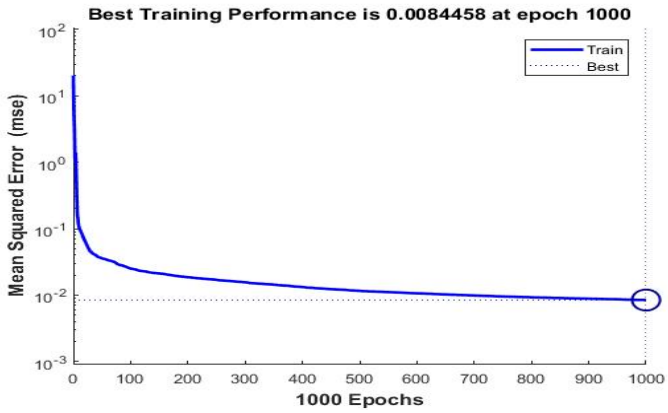

Figure 13. Best MSE for trained deep learning network

It is observed that the minimum gradient of 0.00519505 is achieved at 1000 Epoch and the regression of 0.97903 is obtained demonstrating that the network has reached global minima

point and the weights and bias elements obtained are significant to detect all the 120 objects in the training input and reconstruct the original images with minimum distortions. Figure 13 presents the MSE obtained for reaching the best performance at 1000 Epoch which is found to be 0.0084458. From 800 Epoch the MSE decreases slowly and hence the network reaches its saturation point after 1000 Epoch.

Evaluation of the proposed algorithm is considered by analyzing the model performance of its ability to compute the features that constitute towards salient object detection process.

4. Results & Discussion

In order to identify the significant features of the object present in the input data set the 2D data is converted into 1D data by zig-zag scanning and is plotted as in Figure 14. From the 1D plot obtained for the original image it is observed that the from the origin to the point x1 ($0.65x10^4$) the pixel intensities are between 50 and 130 and from the point x2 ($4.8 x10^4$) till the last point the intensity levels of pixels are in 50 to 130 range. This indicates the background of the object and since the background is gray scale the intensities are closer to zero. From the point x1 to x2 the pixel intensities change and are found to be above 200 level. The intensity levels are mapped into three regions indicated by the points yb1, yb2 and yb.

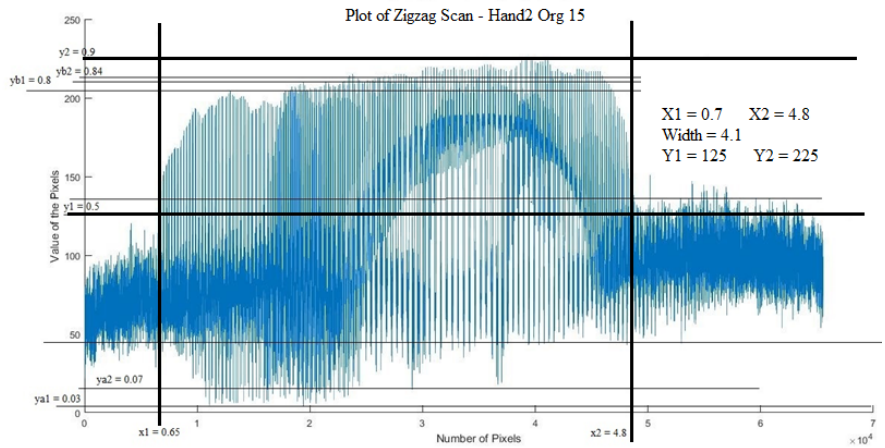

Figure 14. Intensity variation of the object with constant background of original image

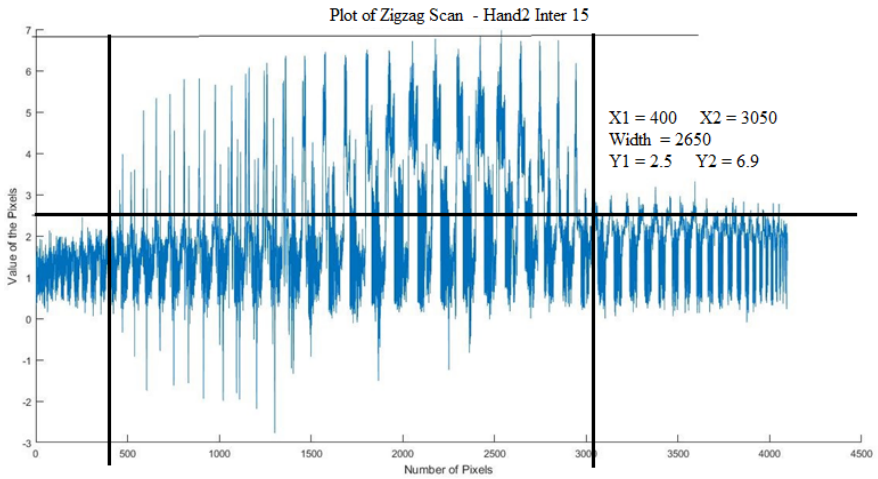

Figure 15. Features representing salient object in the input data

The significant features of the object in the input image exist between 41500 pixels with intensities vary between 125 to 225 units. These features in the original image are captured by the deep learning algorithm as presented in Figure 15. The significant features are present between 400 to 3050 pixel positions and the feature intensities vary between 2.5 to 6.9 units. The original image has 68000 pixels and these features are captured and are present in 2650 pixel positions.

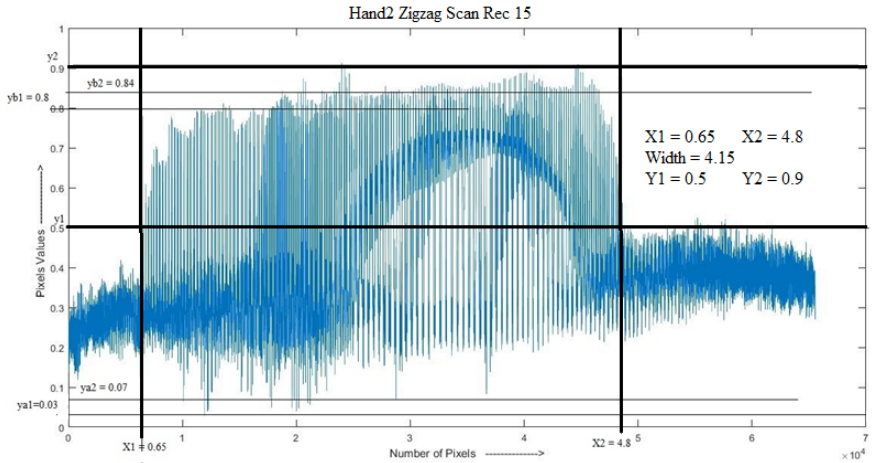

Figure 16. Intensity variations of the object from reconstructed data

From these 2050 pixels that vary between 2.5 to 6.9 units after reconstruction are observed to be present in the same number of pixel positions (shown in Figure 16) as that of input data. The intensities are observed to vary in the range of 0.5 to 0.9 which is of 0.4 differences, and it is observed to be scaled by a factor of 250. Comparing the results of original image and reconstructed image it is observed that the 2050 features captured using the proposed method is optimum in terms salient object detection.

Figure 17 presents the 1D plot of frame 31 and Figure 18, Figure 19 represent the salient object features and reconstructed features.

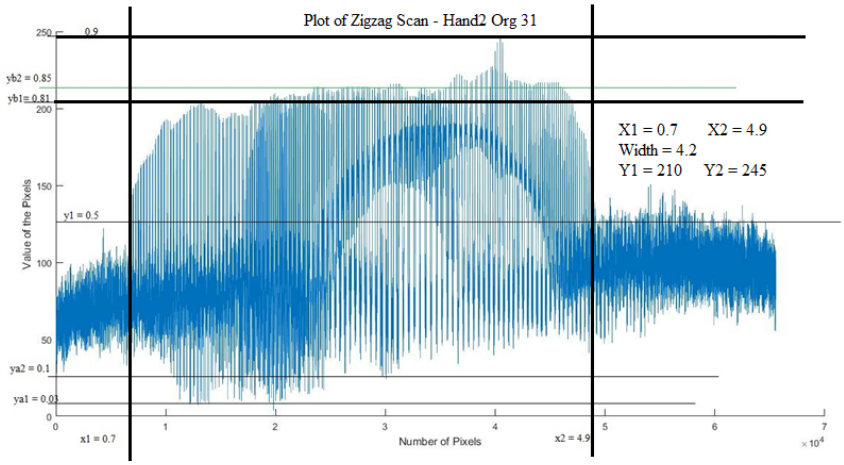

Figure 17. Intensity variations of input data (frame 31)

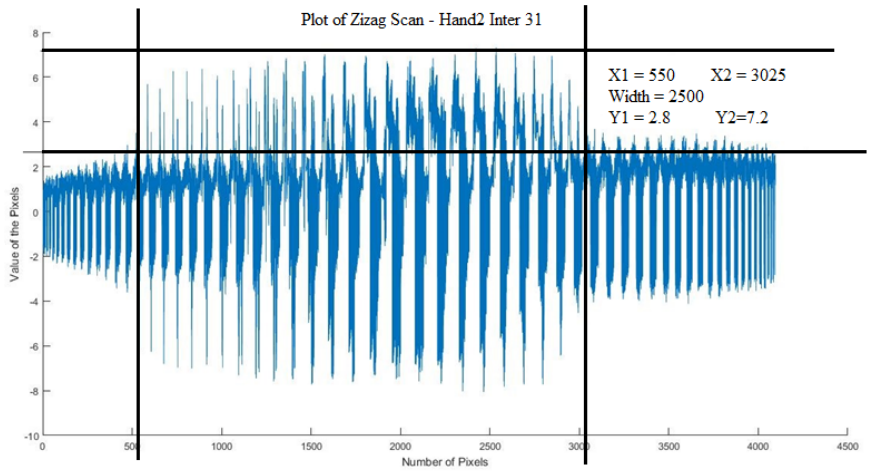

Figure 18. Salient object features captured using deep learning algorithm (frame 31)

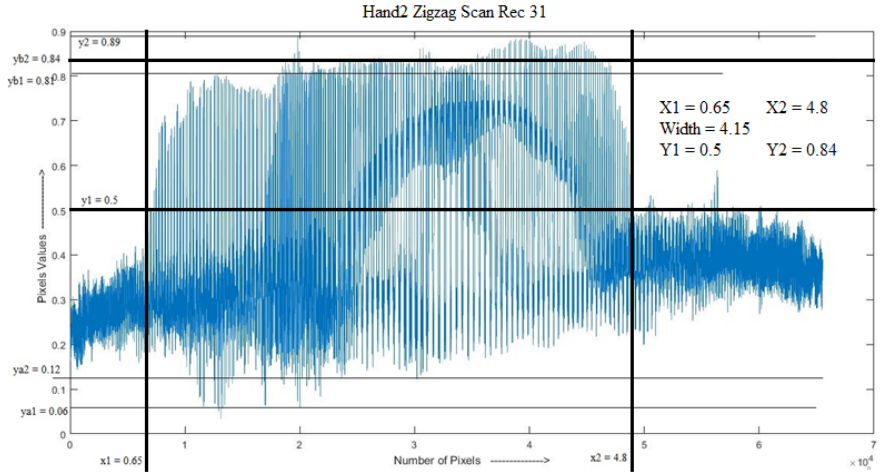

Figure 19. Reconstructed data from salient object features (frame 31)

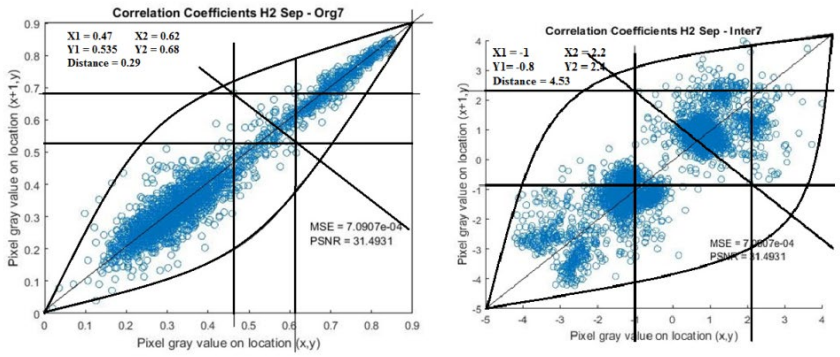

Figure 20. Correlation coefficient plot of (a) original image (b) salient object features

The salient object features are captured in 2500 pixels and is very significant in terms of representing the object features. The reconstructed data is of the same size as that of input image and the pixel intensity is found in the range of 0.5 to 0.9. Figure 20(a) presents the pixel-to-pixel correlation plot that captures the intensity variations of adjacent pixels. The correlation

coefficients are highly interrelated and are placed close to the diagonal axis. Maximum distance between inter-pixel correlations is observed to be 0.29. The original image is encoded to detect the salient features representing the objects in the image and its correlation coefficient plot is presented in Figure 20(b).

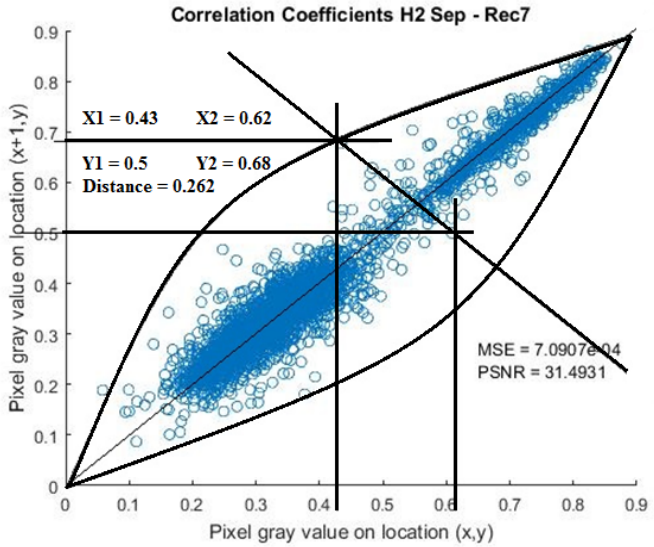

Figure 21. Correlation coefficient plot of reconstructed image

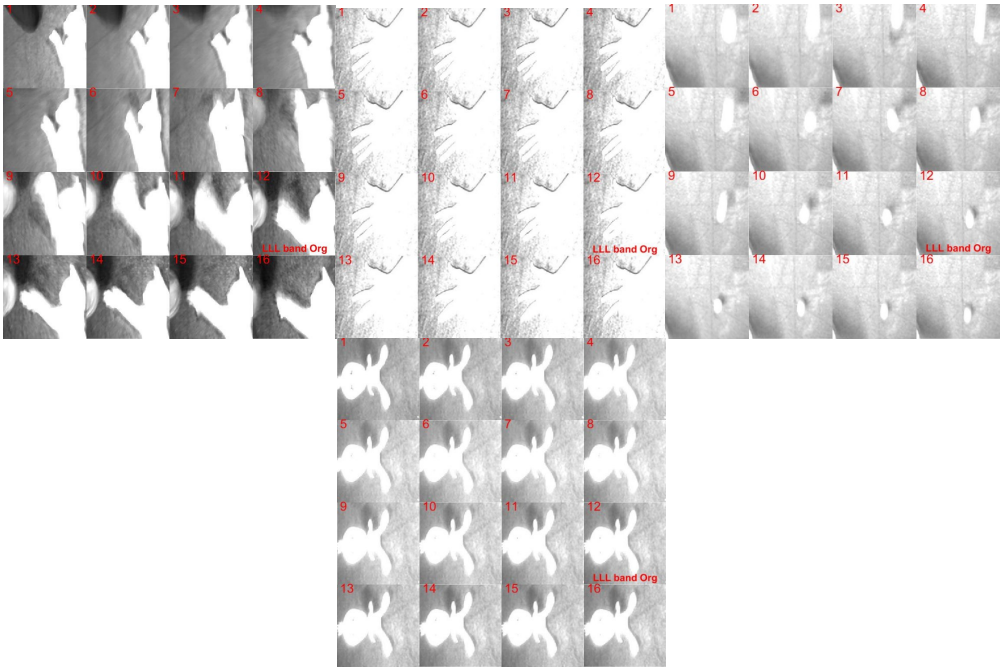

Figure 22. Results of modified SPIHT algorithm (LLL band after quantization)

The correlation coefficients are observed to be concentrated at three different locations along the x-y axis and the maximum deviation with regard to the diagonal axis is observed to be of 4.53. The pixels those are integral part of salient objects are found to be highly correlated as observed at three different locations. Only these pixels capture the information of salient

object that is to be detected from the input image. From these features the input image is reconstructed and the correlation coefficient deviation is observed to be of 2.62 which are similar to the deviations observed in the input image (shown in Figure 21). From the correlation coefficient plot it is observed that the maximum deviations of the features are twice that of the input image. It is also observed that the correlation coefficients are not uniformly distributed across the average axis. Most of the coefficients are concentrated between -3 to 0 and 0.5 to 3 pixel positions along the x-axis.

Figure 22 presents the output of modified SPIHT algorithm that is used to capture the significant features from the multi-level sub bands. The LLL band of the first 16 frames of each of the four different data sets is presented after performing quantization process. From the LLL band it is observed that the intensity or DC component of the objects are distinctly visible in all the LLL bands, the high frequency components that form the edges of the salient objects are not very distinct and are captured in the higher sub bands. In the hand data set, the fingers are clearly visible as there are variations in the intensity level between fingers; the wrist part is completely shaded as the intensity in this region remains constant and is eliminated by the modified SPIHT algorithm. Similarly, the components that are self-similar in all the objects are removed by the quantization process (modified SPIHT algorithm) retaining the significant features that contribute towards salient object detection.

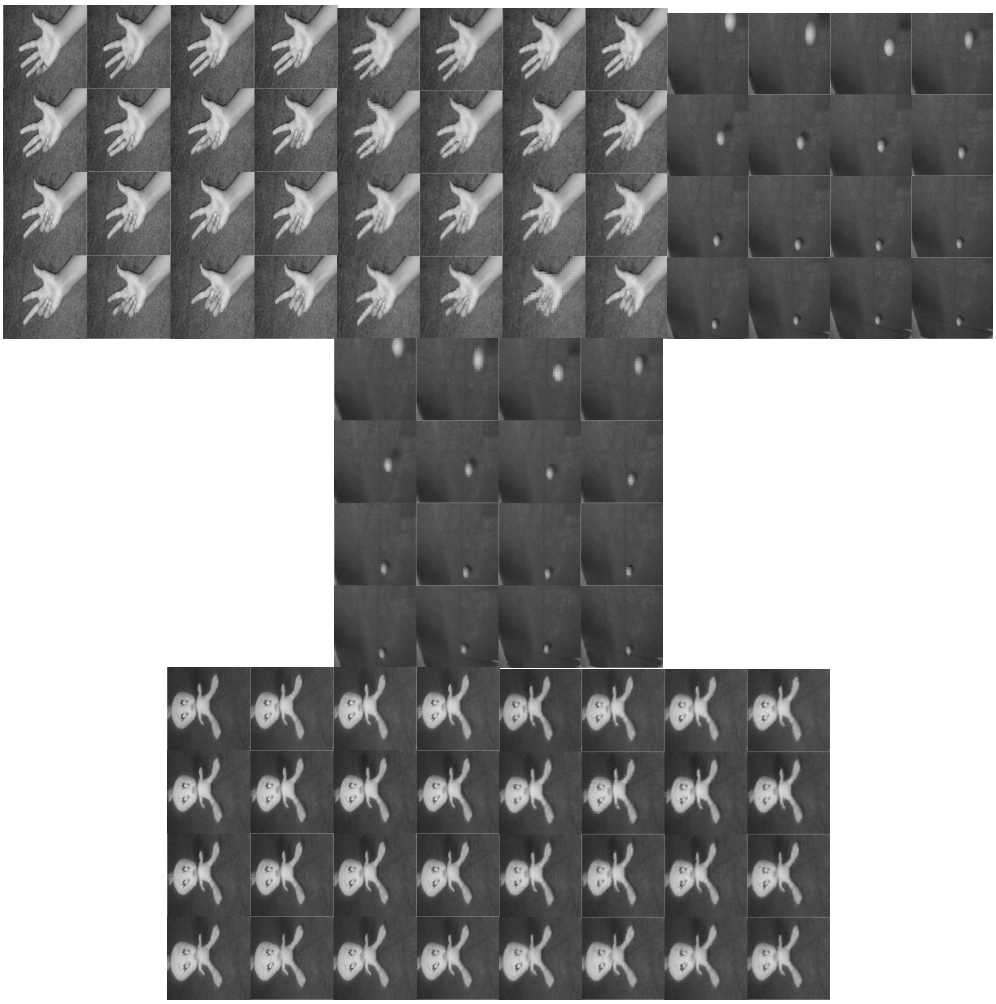

Figure 23. Comparison of input and reconstructed images

Figure 24 presents the input image and the corresponding reconstructed image obtained from the salient object features detected. The input image is pre-processed to standard resolution of N x N x 64 and 3-level 3D DWT is applied to generate the wavelet coefficients. Modified SPIHT algorithm is applied to retain the significant low pass and high pass coefficients and removing the self-similarity components. 2-level 3D inverse DWT is performed to reconstruct the image. Deep learning model of seven layers detects the salient object features. The deep learning model and 1-level 3D inverse DWT processes these features for reconstruction of input data. From the results obtained it is observed that from the salient object features detected it is able to reconstruct the original image with minim distortions. The hand image comprises of vertical, horizontal and diagonal edges that are reconstructed with minimum distortions. The golf ball image that has features in the form of circles, the doll images has features in the form of circles, vertical, horizontal and diagonal edges are also reconstructed with minimum distortions as visible seen results compared.

Table 2. MSE and PSNR results of data sets

	Hand 2		Pink Ball		Torch	
	MSE	PSNR	MSE	PSNR	MSE	PSNR
1	8.7112e-04	30.5992	1.1947e-04	39.2275	2.6177e-04	35.8208
2	0.0015	28.3840	6.9479e-05	41.5815	2.3241e-04	36.3375
3	8.7977e-04	30.5563	1.4517e-04	38.3812	1.9638e-04	37.0690
4	7.0907e-04	31.4931	1.2101e-04	39.1716	2.7113e-04	35.6683
5	8.4082e-04	30.7530	1.4876e-04	38.2750	4.7434e-04	33.2391
6	0.0011	29.5226	1.2791e-04	38.9311	2.9595e-04	35.2878
7	9.8806e-04	30.0522	1.6528e-04	37.8178	2.8161e-04	35.5035
8	8.4308e-04	30.7413	1.2456e-04	39.0460	4.3005e-04	33.6648
9	8.2015e-04	30.8611	1.7345e-04	37.6083	2.9275e-04	35.3350
10	9.8637e-04	30.0596	2.0676e-04	36.8453	1.9526e-04	37.0939
11	8.7009e-04	30.6043	9.0688e-05	40.4245	3.9067e-04	34.0819
12	9.4766e-04	30.2335	8.3700e-05	40.7727	2.7544e-04	35.5998
13	0.0010	29.8676	9.3996e-05	40.2689	2.8467e-04	35.4566
14	9.5248e-04	29.5358	1.3381e-04	38.7351	3.3419e-04	34.7601
15	8.9796e-04	29.4079	2.7093e-04	35.6714	4.8364e-04	33.1548
16	9.9925e-04	30.2114	2.1518e-04	36.6720	3.6087e-04	34.4264
17	8.2650e-04	30.4674	1.0594e-04	39.7492	3.6468e-04	34.3809
18	8.0794e-04	30.0033	1.4711e-04	38.3235	3.8215e-04	34.1777
19	8.0794e-04	30.8276	1.0863e-04	39.6407	3.6384e-04	34.3908
20	0.0012	29.7991	8.6682e-05	40.6207	4.4608e-04	33.5058
21	7.8347e-04	30.9262	1.1984e-04	39.2138	2.7303e-04	35.6380
22	8.8453e-04	29.2398	8.6529e-05	40.6284	3.3244e-04	34.7829
23	0.0011	28.1112	1.3993e-04	38.5410	2.5561e-04	35.9242
24	0.0015	29.2783	8.1874e-05	40.8686	2.9490e-04	35.3033
25	7.0907e-04	31.0598	2.3486e-04	36.2920	2.3049e-04	36.3735
26	8.2650e-04	30.5329	5.2140e-05	42.8283	2.3288e-04	36.3286
27	9.1579e-04	30.3820	5.6313e-05	42.4939	3.1026e-04	35.0828
28	8.6117e-04	30.6491	9.7404e-05	40.1142	3.0180e-04	35.2028
29	7.3333e-04	31.3470	1.8923e-04	37.2300	3.1064e-04	35.0774
30	8.2512e-04	30.8348	1.1172e-04	39.5189	3.3875e-04	34.7012
31	7.8005e-04	31.0788	2.1581e-04	36.6593	4.1536e-04	33.8157
32	9.0089e-04	30.4533	9.3996e-05	40.2689	2.6299e-04	35.8005

Table 2 presents the MSE and PSNR results of three data sets obtained for 31 frames. The input data is encoded and decoded and from each of the frame MSE and PSNR are computed. The PSNR results for all the three data sets are found to be more than 25dB and it is observed that the PSNR variation across all the 32 frames is within +/-2 dB demonstrating that the designed encoder and decoder pair is able to detect salient object features and reconstruct all the frames with motion vectors.

The designed model is capable of reconstructing the salient object from the features without much distortion as visible from the frames presented in Figure 24. The proposed design extracts both spatial and motion vector in the wavelet domain and the deep learning module is trained to detect the salient features that represent the objects. The decoder is trained to reconstruct the original image from optimum number of features. The model designed is able to detect salient object features from more than 120 data sets and is designed for generic object detection. In salient object detection deep learning networks are widely being used as they have demonstrated to be achieving better performance compared to traditional methods. However the complexity of deep learning methods is that there are more than 5 layers of data processing which is addressed by reducing the dimensions of data sets to standards size of 32 x 32. With wavelets being used as feature enhancer there is significant improvement in accuracy of salient object detection. The algorithm proposed in this work uses spatial and motion vectors for salient object detection using deep learning methods.

The reference papers,[25 – 27] are some of the existing methods I, II and III.

Table 3. Comparison with the existing work

	[25]	[26]	[27]	Proposed method
Level of DWT	1, 2,3	1,2 3,4,5,6	1	1, 2, 3
Dimension	2D, 3D	2D	3D	2D, 3D
Wavelet used	db2	db7	--	db2
Window Size	--	64, 128, 256, 512, 1024	--	--
Evaluation metrics	--	--	PSNR, MSE, CR	PSNR, MSE, CR
Compression ratio	--	--	5.73	16
Max PSNR	--	--	47.91	42.82
Min PSNR	--	--	34.56	28.11

The table 3 compares the proposed work with some of the other existing works. The method I [25] uses the DWT at 3 levels, and method II uses the DWT upto 6 levels the method III uses 1 level of DWT, the proposed method uses upto 3 levels of DWT. The method III achieves a compression ratio of 5.73 and the maximum PSNRachieved is 47.91, and the minimum PSNR achieved is 34.56 and these results are achieved on the LANDSAT images dataset. The proposed method achieves a compression ratio of 16 and the maximum PSNR of 42.82 and a Minimum PSNR of 28.11. The Method III uses Huffman coding for encoding and decoding for the compression process of the 3D-DWT coefficients, the proposed method uses SPIHT technique for encoding and decoding of the 3D-DWT coefficients for the process of compression. The method III does not involve the NN for learning the features, whereas the proposed method involves the NN for learning the features. The PSNR in method III is evaluated with original data before encoding and decoding by the Huffman Coding technique and the reconstructed data. The evaluation of PSNR in the proposed method is done with the original data and reconstructed data from the output of trained NN.

5. Conclusion

The 3D DWT model processes 512 x 512 x 64 into m-level sub bands that are processed by the proposed SPIHTL algorithm to identify and remove self-similarity wavelet coefficients and retain the most significant coefficients that constitute towards salient object features. The SPIHTL algorithm is designed to process the multi-band wavelet sub bands considering individual frames. The quantization process eliminates the wavelet coefficients from the SPIHTL ordered list. The quantized frames are rearranged and m-1 level inverse 3D DWT is carried out to reconstruct the image frames. Deep learning algorithm with 2D FFNN, 1D FFNN layers and reordering layers are designed to extract salient object from the video sequences with optimum number of significant features. The decoder module is designed to reconstruct the video sequences from the salient object features. PSNR and MSE measurements are carried out to evaluate the performances of algorithms considering more than 120 different data sets. The developed algorithm is advantageous in identifying the salient object with greater accuracy.

6. Acknowledgment

I thank Dr. K B Raja, Professor in the Department of ECE, UVCE Bangalore, for his valuable inputs and esteemed guidance in bringing out this paper.

7. References

[1]. C. Goldberg, T. Chen, F.-L. Zhang, A. Shamir, and S.-M. Hu, "Data-driven object manipulation in images," Computer Graphics Forum, vol. 31, pp. 265–274, 2012

[2]. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," ACM TOG, vol. 30, no. 6, p. 156, 2011

[3]. U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in CVPR, 2004

[4]. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in CVPR, 2010, pp. 2472–2479

[5]. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," in ECCV Workshop, 2006

[6]. H. Shen, S. Li, C. Zhu, H. Chang, and J. Zhang, "Moving object detection in aerial video based on spatiotemporal saliency," Chinese Journal of Aeronautics, 2013

[7]. Kim, H., Kim, Y., Sim, J. Y., Kim, C. S.: Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing. 24(8), 2552-2564 (2015)*

[8]. Fang, Y., Lin, W., Chen, Z., Tsai, C. M., Lin, C. W.: A video saliency detection model in compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology. 24(1), 27-38 (2014)*

[9]. Liu, Z., Li, J., Ye, L., Sun, G., Shen, L.: Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. *IEEE Transactions on Circuits and Systems for Video Technology. PP(99), 1-1 (2016)*

[10]. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, pages 3431–3440, 2015

[11]. Q. Hou, M.-M. Cheng, X.-W. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In CVPR, 2017

[12]. M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.- M. Hu. Global contrast based salient region detection. *IEEE TPAMI, 37(3):569–582, 2015.*

[13]. H. Liu, L. Zhang, and H. Huang. Web-image driven best views of 3d shapes. The Visual Computer, 2012

[14]. Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018*

[15]. Guanqun Ding, Yuming Fang, Video Saliency Detection by 3D Convolutional Neural Networks
[16]. D. D. N. De Silva, S. Fernando, I. T. S. Piyatilake, and A. V. S. Karunarathne, Wavelet based edge feature enhancement for convolutional neural networks
[17]. Amar, C.B., Jemai, O., et al.: Wavelet networks approach for image compression. ICGST *International Journal on Graphics, Vision and Image Processing pp. 37-45(2007)*
[18]. Said, S., Jemai, O., Hassairi, S., Ejbali, R., Zaied, M., Amar, C.B.: Deep wavelet network for image classification. *In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 000922-000927 (Oct 2016).*
[19]. Szu, H.H., Telfer, B.A., Kadambe, S.L.: Neural network adaptive wavelets for signal representation and classification. Optical Engineering 31(9), 1907{1917 (1992)
[20]. Akhtar, M.S., Qureshi, H.A.: Handwritten digit recognition through wavelet decomposition and wavelet packet decomposition. In: Eighth International Conference on Digital Information Management (ICDIM 2013). pp. 143-148(Sept2013). https://doi.org/10.1109/ICDIM.2013.6693992
[21]. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-SRnet: A wavelet-based CNN for multi-scale face super resolution. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1698-1706 (Oct 2017). https://doi.org/10.1109/ICCV.2017.187
[22]. Williams, T., Li, R.: Advanced image classification using wavelets and convolutional neural networks. *In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 233-239 (Dec 2016)*
[23]. Mohsen, H., El-Dahshan, E.S.A., El-Horbaty, E.S.M., Salem, A.B.M.: Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal (2017)
[24]. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks for texture classification. CoRR abs/1707.07394 (2017)
[25]. Shuihua WANG, Yi CHEN, Yudong ZHANG, Zhengchao DONG, Elizabeth LEE, Preetha PHILLIPS: 3D-DWT Improves Prediction of AD and MCI, *First International Conference on Information Science and Electronic Technology (ISET 2015), pages 60 – 63*
[26]. Jaison Bennet, Chilambuchelvan Arul Ganaprakasam, and Kannan Arputharaj : A Discrete Wavelet Based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis, Hindawi Publishing Corporation Scientific World Journal, Volume 2014, Article ID 195470, 9 pages, http://dx.doi.org/10.1155/2014/195470
[27]. S. Boopathiraja1, P. Kalavathi : A Near Lossless Multispectral Image Compression using 3D-DWT with application to LANDSAT Images, 2018 International Journal of Computer Sciences and Engineering Vol-6, Special Issue-4, May 2018

Suresh Babu D, Research Scholar at UVCE Bangalore, currently working as Associate Professor in the Dept of ECE at MVJCE Bangalore, with 15 years of Teaching experience and 3 years of experience in software Industry. Completed BE Degree in ECE from NIE Mysore affiliated to VTU, completed ME in ECE from UVCE, Bangalore. Areas of interest include – Signal processing, Artificial Intelligence and Machine Learning, and VLSI.

Cyril Prasanna Raj, Professor – ECE, Cambridge Institute of Technology, Bangalore, Director – Sanarys Private Limited Bangalore, Director – ISETILAB Incubator Foundation Bangalore, Council Member – Vision Karnataka Foundation, Total of 25 years of experience of which 8 years of industry experience and 17 years of teaching and research experience, Core competency in deploying & facilitating Problem Solving initiatives including Innovation and product design, Facilitated innovation workshops and generated over 500 ideas in medical and agricultural applications, Incubated and mentored 15 companies and student startups. Field of interest: Electronic Product Development and Services Company working on Medical, Agricultural and Industrial Applications - FPGAs, high speed architectures, wavelets, smart sensors, AI hardware. Education: PhD from Coventry University (UK), M.Tech from KREC (NITK), Surathkal and BE from SJCE Mysore. Research Funding: Successfully completed funded projects worth 7 Cr. from ISRO-RESPOND, DRDO, DST, VGST (CESEM), VGST (CISEE), AICTE (RPS), Atal Tinkering Complete more than 30 minor research projects funded VGST, KSCST, VTU, IE. Core Competencies: DWT architectures over Multicore platform, novel DWT architecture for image compression has been awarded with 4 US patents, DWT architecture on multicore platform work is awarded for patent in South Korea, Japan, US and China. He also has 26 patents, and has commercialized three products. Have published more than 70 journal publications with 300 citations, authored 14 books and is supervising 8 research scholars under VTU (4 graduated from VTU). Supervised more than 300 M.Tech Projects and 40 BE Projects. India's first VLSI design GUI – CYMPLEX and Nanoelectronics Devices Simulator – NANOCYM has been developed by his team at SANARYS Private Limited and is been distributed from his start-up company SANARYS Private Limited in Karnataka, TN, AP and Kerala. IEEE senior member, ISTE LM, Mentor – Atal Tinkering Lab, Brand Ambassador – IIC (MHRD), Evaluator Member – NEAT, AICTE, Has executed 15 consultancy projects for MNCs and DRDOs and delivered more than 50 corporate level training