High Speed and Memory Efficient Structure for Multilevel 2-D DWT

Swati Gupta
Department of Electronics and Communication Engineering, Pranveer Singh Institute of Technology,
Agra - Delhi National Highway - 2, Bhauti, Kanpur 209305 Uttar Pradesh India
swatigupta0891@gmail.com

Abstract **- Wavelet transforms are used in a number of applications. They are applied in different fields such as signal processing, speech and image compression, biometric application and so on. But one of its important applications is image compression, where wavelet transform is preferred over another transform, to compress images. Design of DWT is complex due to a large number of arithmetic operations involved. In this paper, the design strategy of a given architecture for the multilevel 2-D DWT is outlined.**

Keywords: 2-D DWT, Lifting DWT, VLSI, EPI, ADP, Daubechies Filters.

I. INTRODUCTION

THE use of 2-D DWT for image and other signals (lossy compression) is indisputable. Multilevel methods are based on different input transversal pattern. Among these, the most common method used involves the row-column. In this paper, we use the row-column design. The pipeline architecture increases the speed of processing.

A number of architectures have been proposed to provide high speed and area efficient implementation of DWT computation. In this paper, we are using the input image ($N \times N$) size, M multipliers and J levels. In the study of the area, we found all the folded architecture have area complexity 0(N1K). While SIMD (serial input multiple data) has complexities (0NLK), SIMD has complexities of 0(N2K). The time period of folded architecture is approx N2 while the SIMD array is 2JL. An efficient flexibility based architecture for 1- level 2D DWT requires extra buffers. DA based DWT is used to reduce number of multipliers in the polyphase matrix of a wavelet filter decomposed into a sequence of alternating matrix. But these architecture have long critical path; which results in reducing the processing rate of input samples. On the other hand problem of low processing, rate is not acute in architecture that use convolution of low and high pass filtering operations to compute the DWT decomposition of architecture into multilevel DWT to achieve high compression ratio. Being highly memory-intensive, the multilevel 2D DWT is implemented in very large scale integration system to meet the temporal requirement of real- time applications.

Several architectures have therefore been suggested in the last few years for meeting the constraints. Multilevel 2D DWT can be implemented by recursive pyramid algorithm (RPA) but the hardware utilization efficiency (HUE) of RPA based structure is always less than 100% and it also involves complex control circuits (CCC). To overcome this problem, Wu *et al*. [5] suggested a folding & lifting scheme, where multilevel DWT computation is performed level by level using filtering unit and external buffer. Unlike other RPA based design, folded design involves simple control circuitry and it has maximum HUE.

In general, the folded structure based scheme consists of 1-D DWT modules (row-column) processor and storage component device. In the memory, the component consists of a frame memory, transposition memory and temporal memory. The frame memory is basically required to store the (L-L) sub-band. Level-by-level computation of multilevel 2D-DWT transposition memory stores the intermediate results.

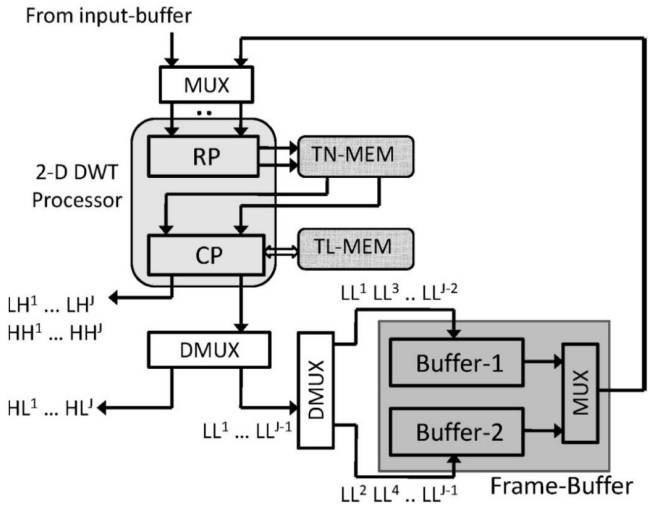

Figure 1. General structure for line based 2-D DWT.

Temporal memory is used by the column processor to store the partial results. Frame memory may either be on-chip or external, while other two are on-chip memories. Transposition memory size depends on mainly the type of data access scheme, adopted to feed the input data, while temporal memory size depends on the number of the partial results.

Sizes of transposition and temporal memory are some multiple of the width of the input signal, while the size of the frame memory is of the order of the image size. On the other hand, complexity of each 1-D module depends on the size of the input signal (image) which is usually very large compared to the size of filter. Complexity of memory component forms the major part of the (overall) complexity of 2-D structure. Cheng *et al.* suggested a parallel (DA) data access scheme to reduce the size of the memory.

Folded structure based on this scheme requires 4N memory words for transposition and temporal memory.

II. THE 2-D DWT ARCHITECTURE

In image or video compression, DWT is basically used for decomposing the signals. If used as a form of a signal using an image, then the image will be decomposed into multilevel 2-D DWT to achieve the high compression ratio of signals. In multilevel 2-D DWT technique, signal compression being highly computation-intensive and memory intensive implementation is in VLSI system for the requirement of real-time applications. Maximum usage is in high data communication and storage through handheld devices. VLSI implementation of 2D DWT is used as a set of incompatible constraints. Many architectures have been suggested for constraint-driven VLSI implementation for 2-D DWT.

Multilevel 2-D DWT technique may also be implemented by RPA, but the hardware utilization efficiency of 2-D DWT is less and also involves complex control circuits. So, for its recovery Wu *et al*. [5] suggested a folded scheme in which multilevel 2D DWT computation is performed using filtering and external buffer. The folded scheme involves simple control circuitry and features maximum hardware utilization efficiency (MHUE) compared to RPA.

Generally, the folded structure consists of the pairing of 1-D DWT modules. Row and column processor and efficient memory storage components consist of frame, transposition, and temporal memory.

Frame memory: this type of memory is basically required to store the (L-L) sub-band of multilevel 2-D DWT.

Transposition memory: this type of memory stores the intermediate values, which result from the row processing, while, temporal memory is the multiple of the input signal. Size of the frame memory depends on the order of the image size or input signal. On the other hand, 1-D module complexity depends on the size of wavelet filter and computation scheme (lifting based) is used to implement the filter. Since the filter size is less and the size of the image signal is very large and complex in memory form and overall complexity of the 2-D structure, Cheng *et al*. suggested a scheme which was based on parallel data access scheme resulting in reduced size of the transposition memory. Folded structure, based on data access scheme requires 4*N* memory words for transposition and temporal memory. The memory requirements in 2-D DWT based lifting structure were more than the folded structures. Meher *et al.* [4] proposed a parallel data scheme for folded structure, based on convolution.

One level 2-D DWT structure requires (K+2) memory which is less in comparison to others lifting based structures, where K is the order of filter (Wavelet filter). Hsia *et al*. suggested the algorithm for 2-D DWT using 5 filters was symmetric, where masks are used for sub-band.

Two-dimensional based wavelet structure does not require the temporal memory, based on the masked algorithm. The 2-D DWT structure requires smaller transposition memory than previous lifting based structure though it involves more additions and multiplications. So, this method was not much better for 9/7 filters. Due to its requirement of large number of multipliers which may not be implemented by shifters, they can only work with 5/3 filters.

Recently work, in a parallel architecture, for 2-D DWT based on on-chip memory, size is 10 N, proceeding with a block of P samples in every motion of cycle. Here, the advantage is that structures keep saving the frame buffers (FB). For achieving the maximum high efficiency of the signal, the minimum block size of multilevel (J level) DWT is $2^{(2j-1)}$. So, the block size of DWT rapidly increases with J level block size, which requires large hardware. Tian *et al.* suggested one level 2D-DWT block based design for high-level implementation. The on-chip memory of the structure varies proportionally to block size. Zhang *et al.* suggested a non-separable approach based on a pipeline architecture for multilevel 2-D DWT to neglect both transposition memory and frame buffer. But the approach of non-separable 2-D DWT is not much popular, as it requires $K/2$ times more memory than the separable 2-D DWT for same throughput rate, where ($K \times K$) is the order number of the 2-D wavelet filter.

Implementation based on convolution has the better solution than the lifting based structure as structure allows row-column folding, if it is in the direct form of convolution based structure without using any temporal memory. The demanding design based on convolution has large transposition memory and more arithmetic components of data than the previous lifting based structure for the equivalent performance.

In this paper, we suggest that we should use the transposition free row-column folded structure for delay of power and area efficient, multilevel 2-D DWT computation. The given structure involves approximately 30% on-chip memory which is 30% less than other lifting based designed structure. So, it

will save the significant area delay and power. As the given structure involves more arithmetic operations or resources than the previous lifting based structure. Designing procedure of given structure is based on parallel data access scheme, which is described in next section.

III. DESIGN STRATEGY

The parallel data access scheme helps in reducing the size of on-chip memory of architecture but at the same time, it increases the complexity of frame buffer. The efficiency of the structure depends on the amount of savings achieved in memory. For determining complexity, first we analyzed the complexity of line based design and the design of folded structure with parallel data access, and then the comparative study of the hardware complexity and the limitations and their usefulness based on parallel data access scheme.

Line scheme based folded structure: This type of folded Structure consists of one processor, temporal memory (TL-MEM) and transposition memory (TN-MEM) and a frame

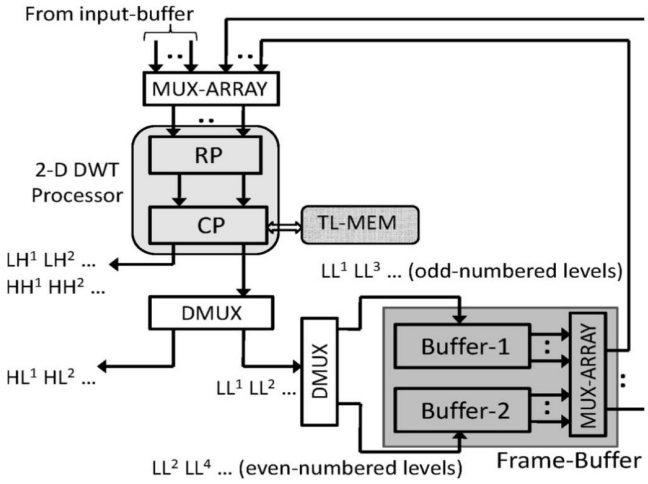

Figure 2. Folded structure for 2-D DWT using parallel data access

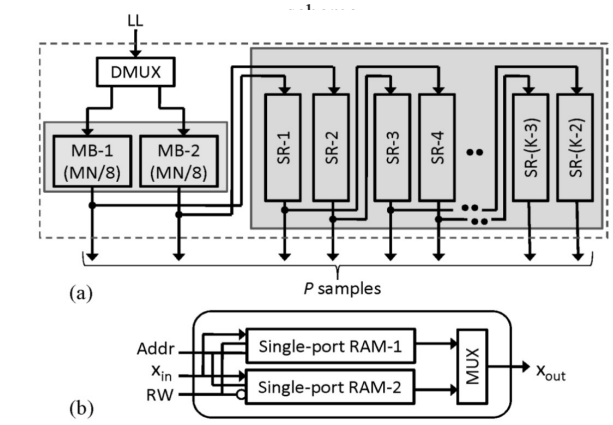

Figure 3. Structure of Buffer-1 and SR using single-port RAM.

The size of upper high and upper low (UL, UH) intermediate matrix forms a pair. In the processing from transposition memory first, column processor receives the (UH and UL) components in the counter clock wise direction and then it generates the four sub-band matrix of signal after decomposition [LL^1, LH^1, HL^1&HH^1] of size of (M/2×N/2) Matrix at each cycle.

Column processor needs a (TL-MEM) Temporal memory for the restoration of temporary and partial results of filtering operation. The size of the transposition and temporal memory are (2Q-1) N/2 and Q′ (N) words; where Q is number of rows of a minimum intermediate components; which may be buffered to column processing initiation. While Q′ is the number of lifting steps of the lifting based filter computation of lifting based DWT structure, where Q = 2, Q′ = *kl*, where *kl* represents the lifting steps of the bi-orthogonal filter (*kl* may also be 2 for 5/3 filter and 4 for 9/7 filter).

Frame buffer will occupy the sub-band of the LL^1 then it will proceed and start receiving data from the frame buffer (FB) soon. In this scheme, two types of frame buffers FB1 and FB2 are basically used as the memory units. Buffer-1 stores the LL sub-band of odd levels and buffer-2 is used for the size of memory MN/4 and MN/16 words respectively.

Parallel data access scheme based on folded structure

Based on parallel data access scheme, folded structure is shown in Figure 2. *P* samples pass through the processor in each cycle of a block. Input blocks of the structure are mainly used from a set of *P* columns of the matrix and the set of adjacent of columns are overlapped by (*P* - 2). Here *P* =3 for DWT structure based on lifting scheme and *P*=*K* for DWT structure based on convolution scheme. Row processor gives the intermediate components column wise and these components directly pass through the column processor by using any transposing memory.

In parallel data, access scheme use of transposition memory will be rejected because it creates complexity to the frame buffer (FB) in the 2-D DWT structure. Hence frame buffer occupies the (LL) sub-bands from the processor in serial form and also feeds those received components block by block back in the processing. In Figure 3, the design of frame buffer(FB) is shown, which mainly consists of two types of memory blocks register (RAM-1 and RAM-2) single port of size (MN/8) words and shifts (P-2) register size (SRs), which basically are used for down-sampling and column overlapping in the structure. Thereafter LL sub-band will split into the odd and even columns by demultiplexers where even and an odd number of columns are buffered in MB-2 and MB-1.

But the simultaneous reorientation of the column from MB-1 and MB-2 are in the same column serial number or order because they were saved. The past column activity is stored in (P-2) SRs. BF-1 used for the data blocks of columns are overlapped by (P-2) columns.

So, the operation of both buffers (FB-1 and FB-2) is same but sometimes it will depend on (MB and SRs), namely MN/32 words and M/4 words respectively. The use of multiplexer (MUX) selects the data of buffers alternately during decomposition levels and SRs may implement by the pair of the single port memory bank (MB) RAM of size M/2 words and one multiplexer.

There are many types of different address signals (Fig. 3b), which are used for the read/write operation; and control circuits for the frame buffer (FB).

Analysis of structure based on its complexity: Structure based on parallel data access scheme first put out the transposition memory (TN-MEM) but it increases the memory size of the frame buffer (FB). For image size of (512×512), we realize memory saving by the parallel data access scheme over the line based scheme for different types of wavelet filters and the values which are plotted.

Thus data access scheme based on parallel filters offers more memory saving in the case of convolution based DWT than lifting based DWT structure. In the case of Daubechies-4 filters, it will increase with the order of the filter. Temporal memory occupies large lifting based folded structure. Lifting structure based on chip memory with parallel data access scheme includes line buffers. On the other hand, the structure based on convolution scheme with parallel data access scheme does not occupy the transposition memory and temporal memory. In terms of advantages, the design of multilevel 2-D DWT structure is memory efficient, low in terms of computation time without frame buffer (FB). Since the saving of memory achieved by parallel data access scheme (PDAS) can reduce the overhead cost of the frame buffer (FB). In this paper, we have also observed that if we do not want to use FB then we can eliminate it by the multilevel DWT in a pipeline structure.

Observation on proposed design strategy

As compared to convolution based approach, lifting based scheme is better because this approach features lower complexity of arithmetic. After analysis of convolution and lifting based design, while lifting based design have large complexity in comparison to convolution based scheme with proper scheduling of multilevel 2-D DWT decomposition, so the memory saving capacity of convolution based structure is higher than the saving of arithmetic components.

By using parallel data access, memory complexity can be reduced significantly; it can also use the maximum space of memory requirement.

TABLE 1 - MINIMUM INPUT BLOCK SIZE FOR DIFFERENT DWT LEVELS

DWT levels *(J)*	**Input block size** *(P)*
1	1
2	4
3	16
4	64
5	256

If down sampled filters are used in the proposed structure then, the block sizes of structure can be used efficiently for 2-D DWT levels. By using parallel data access scheme in the structure, one could achieve the maximum use HUE. Input block size (P) of the structure depends on the availability of resources. If blocks-size is not sufficient for designing of higher level 2-D DWT structure, the line based scanning with RPA based computation is considered.

Thus a parallel and pipeline-based architecture for three levels 2-D DWT given architecture is suitable for both types of filters Daubechies as well as biorthogonal wavelet filters.

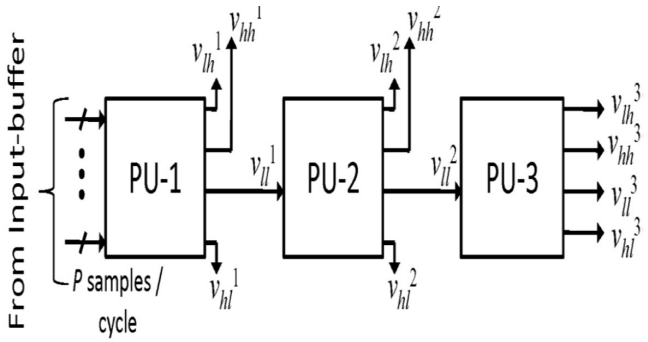

Figure 4. Computation structure for three level 2-D DWT.

IV. COMPARISON BASED ON COMPLEXITY AND PERFORMANCE

Comparison based on theoretical complexities estimation: In the proposed structure, hardware complexity as well as previous convolution based structure and lifting based structures, complexity in term of ACT and minimum Clock period, address, and external memory FB words, on-chip storage words is listed in Table 2 [10].

TABLE 2 -- COMPARISON OF ARITHMETIC, MEMORY AND TIME COMPLEXITIES OF THE PROPOSED STRUCTURE, AND THE EXISTING STRUCTURES FOR 3-LEVEL 2-D DWT OF AN IMAGE OF SIZE (512 × 512) WITH DAUBECHIES 4-TAP FILTERS AND 9/7-BIORTHOGONAL FILTERS

Structure	MULT	ADD	MEM words	CT (in Ta)
Huang [6]	16	16	83724	344064
Maher [8]	16	12	83462	344064
Cheng [9] (8parallel)	96	240	6752	65535
Proposed (Daub-4)	168	126	1050	32768
Xiong [11]	18	32	84742	344064
Lai [10]	10	16	84780	516096
Cheng [9] (2parallel)	24	76	16598	436905
Mohanty [13] (P=16)	99	176	5854	65536
Tian [14](P=16)	96	128	88448	86016
Proposed (9/7filter)	189	294	3131	32768

For comparison, on-chip storage represents the sum of registers, SRs by the core, RAM words.

In table 2, a structure needs small in number on-chip memory than the pre-existing folded lifting based structure [11]. So, the presented structure is more efficient in use. Two arithmetic components (adder & multiplier) are required than those [10] and it gives double throughput rate. Still, it includes low FB and on-chip memory, compared to the proposed structure. This requires (K0/K1) times more multiplier and address ≈ (3k1M/4k1N) times more on chip (storage words) device & (16/3) less automatic computation time (ACT). Where K0 = K1 + K2, hence in comparison to the parallel structure given proposed structure needs fewer multipliers, adders, computation time, less on-chip storage after less throughput rate.

V. SIMULATION RESULTS

Simulation is done with synchronization of proposed architecture based on Daub-4 filters with convolution and lifting based design of three level 2-D DWT by synopsis design compiler using TSMC CMOS library. Since we have synthesized core of [6], [8]& [10], the frame buffer is used in the architecture as an external to the chip. For the design of [13] and for proposed design, we considered Q=16. For the design of architecture, we have used 8-bit input pixel and 12-bit intermediate signals and the size of an image (512×512) for all designing purpose.

We used design-ware building blocks library for Wallace tree based booth multiplier. A net-list file is used in synopsis IC compiler. For routing the area, power and time by IC compiler is given in table 3. For implementation of the frame buffer, DRAM is used after using approximately 138671.9 μm^2. In the 90 nm process for 1-MB DRAM, we assumed the approximate area of FB for comparison, Area delay product (ADP) & ACT clock cycles required by the design to complete the multilevel 2D-DWT. After the theoretical study synthesis, results are given in Table 3.

[ADP= area× ACT× data arrival time (DAT)

Here, area represents the frame buffer (FB) area or core area (CA).]

The structure based on lifting has lowest DAT (data arrival time).

TABLE 3 -- COMPARISON OF SYNTHESIS RESULTS OF THE PROPOSED STRUCTURES AND THE EXISTING CONVOLUTION- AND LIFTING-BASED STRUCTURES FOR DWT LEVELS *J* = 3 AND IMAGE SIZE (512 × 512) (TSMC 90 NM CMOS TECHNOLOGY LIBRARY, POWER ESTIMATED AT 20 MHZ FREQUENCY)

Structures	**Block** 1	**DAT**(ns)20	**Core area** ($\mu m)^2$	**Core power** (mW)	**FB area** (μm^2)	**ADP** (μm^2.s)	**Core-EPI**(μJ)
Structure of	2	20.42	810395.12	6.3451	136320.02	3325.7	54.57
Structure of	2	17.73	64381.68	0.8284	138376.80	618.43	7.12
Proposed Daub-4	16	21.6	1042372.16	7.2021	0	362.5	5.87
Structure of	2	15.8	944305.45	6.997	137598.02	2940.72	60.18
Structure of	16	45.58	3104371.05	22.5874	0	2318.29	18.50
Structure of	16	42.66	3241550.68	24.4466	136320.02	3098.72	26.28
Proposed (9/7)	16	25.42	2139397.29	15.2605	0	891.01	12.50

Using Daub-4 filters, the proposed structure includes some more core area compared to before and less automatic computation time (ACT). But it will not include frame buffer (FB) external memory. In proposed structure of 2D-DWT using (9/7) filters, it involves approximately 2.26 times more core area and 10.5 times less (ACT). While the proposed and parallel structure of 2D-DWT used same computation time it includes 1.51 times less core area as well as less ACT. In the observation of proposed structure it used 3.3, 3.4 and 2.6 times less ADP (area-delay product) than [10], [13] and [14] respectively.

Energy Consumption of Architecture : Energy consumption of architecture can be calculated by the power consumption by core × computation time. For Daub-4 filters, energy consumption is 9.29 and 1.21 times less EPI (energy per image) than other, which has been shown in the table . In the proposed structure for (9/7) filters it consumes 1.4×8 and 2.1 times less EPI. Since, for calculation of EPI, the EPI of structure would be higher than the assumed value.

V. CONCLUSION

Memory complexity is a big issue for the realization of 2D-DWT in VLSI system. So, focusing on the suggestion a memory-centric design, based on memory a convolution based architecture for the computation time have been derived from three level 2-D DWT architecture based on orthogonal as well as Daub-4 filters. In proposed structure, it will not use the frame buffer (FB) but it occupies the line buffers (LB) of size 3(K-2) M/4 which is not dependent on throughput rate. So, this is the advantages of architecture when it is designed for higher throughput rate. In this paper proposed structure for (9/7), filters for image size (512×512) gives the low complexity of area and less computation time. Result based on ASIC shows that the proposed structure could be used for the energy implementation of multilevel 2-D DWT as well as area delay using Daubechies and biorthogonal filters for high performance of processing of image application.

VI. REFERENCES

[1] Yiqi Zhou Xiaodong Xu, "Efficient FPGA Implementationof 2-D DWT for 9/7 Float Wavelet Filter," *IEEE*, september 2009.

[2] B. K. Mohanty and P. K. Meher, "Memory-Efficient High-Speed Convolution-Based Generic Structure for Multilevel 2-D DWT," *IEEE Trans. Circuits and Systems*, Volume 23, Number 2, June 2012, pp. 353-363.

[3] A Memory-Efficient Stripe Based Architecture for 2D Discrete Wavelet Transform, *IJIRSET*, Volume 3, Number 1, February 2014, pp. 1564-1570.

[4] Pramod Kumar Meher and Basant Kumar Mohanty, "Memory-Efficient High-Speed Convolution-based," *IEEE Trans Circuits and Sysytems,* Volume 23, Number. 2, February 2013, pp. 353-363.

[5] L. Wu, Y.-H. Tan and J.-W. Tian X. Tian, ""Efficient multi-input/multimultioutputVLSI architecture for 2-D lifting-based discrete wavelet transform", *IEEE Trans.*, Volume 60, Number 8, August 2011, pp. 1207–1211.

[6] C.-T. Huang, C.-Y. Cheng, C.-Jr. Lian and L.-G. Chen, C.-C. Cheng, "On-chip memory optimization scheme for VLSI implementation of line-based 2-D discrete wavelet transform", *IEEE Trans.*, Volume 17, Number 7, July 2007, pp. 814-822.

[7] B. K. Mohanty and P. K. Meher, "Memory-efficient modular VLSI architecture for high-throughput and low-latency implementation of multilevel lifting 2-D DWT," *IEEE Trans.*, Volume 59, Number 5, May 2011, pp. 2072–2084.

[8] http://www.i2r.a-star.edu.sg/publication/energy-efficient-high-speed-convolution-based-generic-structure-multilevel-2-d-dwt.

[9] J.-M. Guo and J.-S. Chiang, C.-H. Hsia, "Improved low-complexity algorithm for 2-D integer lifting-based discrete wavelet transform using symmectric mask-based scheme," *IEEE Trans.* Volume 19, Number 8, August 2009, pp. 1202–1208.

[10] Lien Fei Chen, Yui-Chin Shin and Yeong-kang Lai, "A High-Performance and Memory-Efficient VLSI Architecture with Parallel Scanning Method for 2-D Lifting-Based Discrete Wavelet Transform," *IEEE Trans*, pp. 400-407, June 2009.

Swati Gupta was born in Etawah, Uttar Pradesh. She received B.Tech degree in Electronics and Communication Engineering from Gautam Buddh Technical University, Lucknow in 2013. Currently pursuing M.Tech from Pranveer Singh Institute of Technology, Kanpur.

Her research interests include image and digital signal processing algorithms and VLSI architecture development.

Worked towards Master's degree in image processing with thesis topic: "Efficient Design of Architecture for Image Compression using 2-D DWT". Attended national conferences, faculty development program and research and development programs in Kanpur and Ghaziabad. She is sincere, dedicated, and fully passionate with her work.