Chapter 6

Conclusions and Future work

The researches on video coding standards have been evolving since 1990s to satisfy the growing demand of video. Thus, HEVC was introduced in early 2013 to provide a coding gain of $40 - 50\%$ compared to its predecessor, H.264/AVC [11]. However, the complexity of the HEVC codec is very high. Consequently, resource and energy requirements of the encoding and video playback devices are very high. On the contrary, the popularity of portable battery operated devices is driven by area and power constrains. Hence, complexity reduction without significant coding loss is essential for HEVC. A study carried out on HEVC profile, shows that a significant amount of time is spent on transform and quantization function during video encoding process. According to this study, about a quarter of the total encoding time is spent in the transform and quantization in the AI encoder configuration. Therefore, efficient forward and inverse transform architectures are essential. Besides this, an effective approximation of these architectures would be useful to reduce area and power requirements. Hence, an attempt is made in this research work to reduce the hardware complexity of the transform block in HEVC with minimum compromise in the coding efficiency. Few transform architectures have been proposed to satisfy low complexity, low power and high throughput requirements of the HEVC core. The accuracy and the video encoding performance of all the proposed architectures are measured using HEVC reference software model [12] of version 16.15 (i.e., HM-16.15). All the hardware architectures are described using Verilog hardware description language. Synopsys Design Compiler and Xilinx Vivado Design Suite version 2016.2 are used for hardware implementations.

The main objective of this thesis is to improve the performance of the core transform in HEVC standard. The transform core used in HEVC contributes a large amount to the hardware complexity which is the major concern for the HEVC to be implemented in real-time systems. Some architectures have been proposed in this thesis which increase the performance either by reducing hardware complexity or by increasing the throughput. Following are some conclusions based on this study:

- A 32-point DCT architecture for HEVC with a new set of fixed point coefficients is proposed in this thesis. This coefficient set approximately holds all the DCT properties with minimal error and it is comparable to that of the HEVC core transform matrix. But, the transform can be realized by integer multiplication and shift operations. Therefore, the drifting error can be reduced by standardizing this integer numbers and shift amounts both for HEVC encoder and decoder. Intermediate data length and complexity of the proposed architecture is less. This results optimized hardware architecture for HEVC core transform. Additionally, it is proven that a transpose memory of 15-bit data depth is enough to calculate 2D DCT of 9-bit residual data. When the proposed 1D DCT architecture is implemented on Virtex-7 XC7V2000T FPGA, it operates at 122.6 MHz. CMOS 90 nm ASIC implementation of this architecture requires 88.6K logic gates to produce constant throughput irrespective of DCT size and it operates at 256.4 MHz.

- New HEVC compliance WHT based generalized model for N-point DCT and IDCT architectures are also proposed in this thesis. The hardware of proposed architectures can be shared between prediction stage and transform kernel of the HEVC encoder as well as decoder thereby reducing the overall hardware -cost and -area.

- The trade-off between coding accuracy and computational complexity has been studied in this thesis. It has been shown that perceptual quality does not affect significantly with a slight modification in the higher order DCT/IDCT coefficients. So, an approximation technique is proposed in this thesis which reduces computational complexity. The proposed approximation algorithm replaces the rotation unit by a butterfly structure such that orthogonality and normality errors are minimized. Therefore, rotations are inessential with this technique and resultant architecture becomes multiplier-less.

- The effect of the proposed approximation scheme on coding performance has been studied on different size of DCTs. Based on it, four different DCT and an IDCT architectures are implemented which maintain different trade-offs between coding accuracy and hardware complexity. It has been observed that the proposed method is more accurate as compared to the other approximation techniques and produce better results. The proximity measures show that the maximum mean square error attained using any of the proposed architectures is 4.06 with the minimum coding gain 8.95. The proposed 1D DCT architecture requires 111.23 K logic gates at 250 MHz operating frequency without any approximation on CMOS 90 nm ASIC platform. However, with the approximation algorithm, 43% to 82% savings in area-delay product can be achieved depending on the accuracy required. When implemented on Virtex-7 FPGA, the proposed approximated architecture reduces power consumption by 70% as compared to that of the HEVC reference architecture to maintain a reasonable trade-off between accuracy and complexity.

- Transform size in HEVC varies from 4×4 to 32×32. But, a single core is used to compute these different sizes of transforms. Consequently, throughput of the core reduces during lower size transform operations. It has been observed that for any encoder configuration lower order of DCTs are performed very often than that of the higher order irrespective of the video resolution. Study performed in this thesis shows that on an average more than 50% of times 4-point DCT is performed, whereas 32-point DCT is performed less than 2% of the time. Therefore, most of the time majority of the transform core hardware resources remain idle and throughput suffers. So, a DCT architecture based on hardware sharing methodology is presented in this thesis and high throughput with the minimum number of logic resources is achieved.

- Since lower size DCT blocks are frequently used in the video coding process, the throughput of the core transform depends on how quickly lower size DCT is processed. A constant throughput integer DCT architecture proposed in this thesis improves the RDO speed and maintains almost constant processing time irrespective of quadtree depth. The proposed 1D DCT architecture can process 32 samples/clock irrespective of the transform size. Further, the proposed modified 2D DCT architecture requires $2N+1$ clock cycles to process $\frac{32}{N}$ blocks of $N \times N$ pixels. Depending on the size of transform to be performed, multiple sets of DCT constants are produced using reconfigurable MCM design. However, adopting A-operation the critical path in the reconfigurable MCM is restricted only to two adders which further improves the operating speed of the architecture. The proposed 1D DCT architecture operates at 104.6 MHz when implemented on Virtex-7 FPGA and it can process 3347.6 M samples per second. CMOS 90 nm ASIC implementation of the same design requires 79.2K logic gates and it operates at 157 MHz. The proposed folded 2D DCT architecture requires 87K gates and its throughput is 2.3G pixels per second. The maximum frequency of operation of this architecture is 146.4 MHz and power consumption is 28 mW. Therefore, the proposed 2D DCT architecture can process at most 185 frames and 46 frames of 4K and 8K UHD video, respectively.

Finally, the contribution of the thesis can be listed as follows:

- An alternative to the integer DCT is proposed in this thesis for HEVC. The proposed method uses real-valued DCT coefficients which reduces the dynamic range of the transform architecture.
- New data flow models of WHT based DCT and IDCT architectures are proposed. The hardware of the proposed models can be shared with the prediction unit in HEVC.
- An algorithm is proposed for approximating DCT as well as IDCT architectures in HEVC. The proposed algorithm significantly reduces the hardware complexity.
- The effect of the proposed approximation algorithm on different size of DCT has been analyzed to demonstrate the trade-off between the coding performance and hardware cost.
- An approximate higher order WHT and on its basis pruning of high frequency DCT coefficients is proposed. The proposed pruning method has negligible impact on the coding efficiency, but reduces hardware complexity significantly.

- A novel hardware sharing method is proposed which maintains the constant throughput irrespective of DCT size. The proposed hardware sharing method can improve the speed of RDO method in HEVC.
- The coding performance of all the proposed architectures is assessed using reference software HM-16.15. All the proposed architectures comply with the requirements of HEVC standard.

The video encoding technology today has undergone tremendous changes over the last few decades. The popularity of video applications is increasing, such that video data occupies a large percentage of mobile data traffic nowadays. Hence, the video compression has become a crucial element in the media distribution chain. This thesis has focussed only on transform block which is a sub-module of video compression unit. The research work carried out in the thesis is able to satisfy the demands of video applications. However, these demands are changing rapidly. Video coding technology is facing new challenges with the advancement in mobile technologies, the popularity of high resolution video contents and video sharing applications. To fulfill those, further research is definitely needed.

This thesis presents few architectures for transform coding in HEVC to reduce the complexities incurred in the encoder and decoder. Still, there is huge scope to reduce the complexity of transform architecture in HEVC. Those need to be explored further. A detailed study of the accuracy and complexity requirements targeting resource constrained mobile playback devices is necessary.

The transform block in HEVC uses DCT for energy compaction. The compression efficiency can be further investigated with more higher order DCTs. Exploration of different types of DCT, e.g., 3D DCT, Directional DCT (DDCT) is ongoing to further enhance the performance. Investigation of some other alternative transforms instead of DCT in HEVC reference and its hardware implementations can be carried out.

In this thesis, the coding performance assessment is performed with the HEVC reference software version HM-16.15. However, the reference software also supports other extensions of HEVC, i.e., 3D-HEVC, MV-HEVC, SHVC, and RExt extension [120]. It is important to see the effects of the proposed transform models in these HEVC extensions. It would also be highly useful to conduct the experiments using more test video sequences of HD quality and beyond, as well as test suites suitable for the evaluation of range extension, scalable coding, multiview, and 3-D applications.

This thesis has concentrated on the analysis of the transform operations in video coding and the other aspects such as motion estimation/compensation, filtering, human visual system based lossy coding are not considered in this work. However, motion estimation and filtering also require intensive computations and the researches in this field has great importance for future video coding. The future work could focus on a new RDO scheme, more efficient entropy coding techniques, reducing the memory requirements, incorporating more efficient intra- and inter- prediction mechanisms. Moreover, the HVS based lossy coding is expected to provide better subjective video quality to the end-users. Thus, potential future work could provide more analysis on the aspects of perceptual quality.

Finally, transform architecture is a mere sub-block of the complete codec. Effective performance of the entire codec depends on the performance of all other sub-blocks. Therefore, it is imperative to improve their performance too! One can start working on the rest of the blocks in HEVC and try to improve its performance in future. However, it is only possible with research collaborations, team work and good source of research funding.