VLSI architectures for the Steerable-Discrete-Cosine-Transform (SDCT)

Luigi Sole, Riccardo Peloso, Maurizio Capra, Massimo Ruo Roch, Guido Masera, and Maurizio Martina

Politecnico di Torino

Abstract. Since frame resolution of modern video streams is rapidly growing, the need for more complex and efficient video compression methods arises. H.265/HEVC represents the state of the art in video coding standard. Its architecture is however not completely standardized, as many parts are only described at software level to allow the designer to implement new compression techniques. This paper presents an innovative hardware architecture for the Steerable Discrete Cosine Transform (SDCT), which has been recently embedded into the HEVC standard, providing better compression ratios. Such technique exploits directional DCT using basis having different orientation angles, leading to a sparser representation which translates to an improved coding efficiency. The final design is able to work at a frequency of 188 MHz, reaching a throughput of 3.00 GSample/s. In particular, this architecture supports 8k UltraHigh Definition (UHD) (7680x4320) with a frame rate of 60 Hz, which is one of the best resolutions supported by HEVC.

Keywords: Video Coding, Discrete Cosine Transform, Directional Transform, VLSI

1 Introduction

In recent years, a large effort has been devoted to the field of video compression to cope with the increasing demand of high resolution multimedia contents. The latest standard proposed by ITU-T and ISO/IEC groups is the H.265/HEVC compression algorithm [8]. It extensively employs inter-frame and intra-frame prediction to exploit the temporal and the spatial redundancies present in video streams. H.265/HEVC requires computational load to detect and process intra mode, so many efforts have been done in order to lower the complexity [6] of the detection phase. The difference between the predicted block and the actual block of pixels is called *residual block* and it is lossly coded taking advantage of transforms (Discrete Sine Transform, DST, and Discrete Cosine Transform, DCT) and quantization. While the DST is used only for the smallest block size, namely 4x4 pixels, the DCT is used for all the other sizes, typically up to 32x32. Chen et al. [1] has shown how to reduce the complexity of the Integer Cosine Transform enabling solution up to 64x64. Since DCT is increasing in complexity and computational load, faster and low-power architectural solutions such as [9] and [7] are required. Recently, G. Fracastoro et al. [3] proposed a

directional DCT, called Steerable DCT (SDCT), which is better suited than DCT to compress directional data. The SDCT is based on the work of B. Zeng et al. [10] and makes possible to divide the directional cosine transform into a traditional DCT followed by a geometrical rotation. The kernels used for the SDCT are different from the DCT ones as they depend on the steering angle, with the limit case of 0 degrees rotation for which the SDCT coincides with the DCT. This paper presents a low power hardware accelerator for SDCT able to reach the throughput required by HEVC for the 8k UltraHigh Definition of 7680x4320 pixels. At first the architecture is analysed in Section 2 and then the Section 3 will present the obtained results for the basic SDCT accelerator and some implementations stemming from it.

2 Architectural implementation

While the 2D-DCT employed in HEVC is an inherently separable operation, the SDCT must be computed all at once. The complexity of a transform that is not separable is far greater than a separable one, so this may be a big drawback for the implementation. However, the complexity can be decreased drastically by splitting the SDCT in two parts, namely a separable 2D DCT followed by some rotations, and then by computing the separable transform before applying rotations, as reported in [4]:

$$\tilde{\boldsymbol{x}} = T(\boldsymbol{\theta})\boldsymbol{x} = R(\boldsymbol{\theta})T\boldsymbol{x} = R(\boldsymbol{\theta})\hat{\boldsymbol{x}} \tag{1}$$

where $\boldsymbol{x}$ are the input samples, $\hat{\boldsymbol{x}}$ are the results obtained by applying the T transform matrix, $R(\boldsymbol{\theta})$ is the rotation matrix, while $\tilde{\boldsymbol{x}}$ is the result of the SDCT. The SDCT can be thus implemented as a DCT followed by a steering transformation. The DCT part can be implemented as suggested in the literature, for example using a folded architecture [5], and then applying rotations when all the samples returned by the 2D-DCT are available. This means that the steering part of the architecture, which handles the rotations, has to work faster than the DCT. This issue has been addressed in this work and one of the possible solution is to define two clock regimes, one for the 2D-DCT and one, faster, for the steering part, in order to comply with the throughput offered by the 2D-DCT transform block. A FIFO memory between the two parts acts as a buffer memory. The whole structure is depicted in Figure 1. The 2D-DCT block is based on the architecture proposed in [5] by Meher et al., which is very efficient, especially in the folded fashion, and scalable to transforms of size 4, 8, 16 and 32. The steerable part is shown in Figure 2. It is composed by an input memory (IM), an output memory (OM) and the lifting blocks that perform the rotation [2]. Some multiplexers are used to bypass the lifting blocks for the case of no rotation, returning directly the result given by the DCT. The IM is required also to reorder the samples as the steering process is computed on the custom zig-zag order, given in Figure 3, that is different from the classic zig-zag ordering, as the vectors are rotated in pairs with respect to the diagonal elements. Rotation by

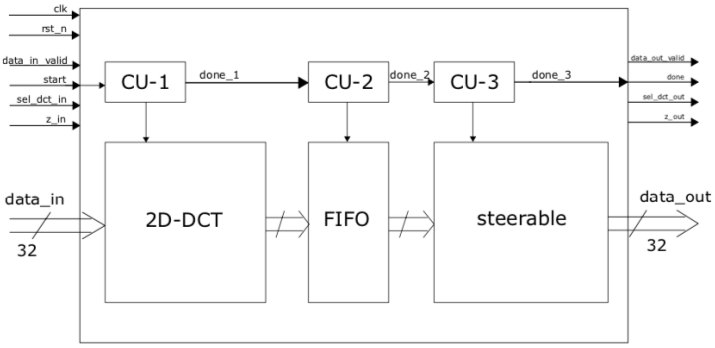

Fig. 1. Whole SDCT structure

lifting scheme:

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 1 & \frac{1-\cos\theta}{\sin\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\sin\theta & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1-\cos\theta}{\sin\theta} \\ 0 & 1 \end{pmatrix} \tag{2}$$

The rotation matrix is decomposed in the multiplication of other three rotation matrices, in such a way the resulting structure, shown in Figure 4, presents a lower complexity. Indeed, this implementation requires only three multipliers, while the original rotation matrix would need four multipliers to achieve the same result. In order to further simplify the architecture, the multiplication for P and U coefficients from Equation 2

$$P = \frac{1-\cos\theta}{\sin\theta} \tag{3}$$

$$U = -\sin\theta \tag{4}$$

in Figure 4 is implemented as shift and add, as the number of possible rotation angles have been fixed to 8 (from 0, no rotation, to 7), as reported as optimum in [4] by M. Masera et al. The steerable block thus introduces $2 \times N$ clock cycles of latency for the reordering stage plus 4 clock cycles due to the internal pipeline. Therefore, in the event that all the SDCT have a length N=32, the latency is equal to 68 clock cycles, which corresponds to the worst case.

2.1 Reduced SDCT architectures

The unit presented so far is able to compute SDCT of lengths 4, 8, 16 and 32. This type of structure has been designed to be implemented inside the HEVC standard. Anyway, this algorithm could be also used for video compression standards with lower constraints and for image compression standard, such as JPEG. Therefore, two reduced SDCT unit have also been developed. The first is able to compute SDCT of length 4, 8 and 16, named SDCT-16, while the second is capable of computing SDCT of length 4 and 8, named SDCT-8. These two units have a reduced throughput of 50% and 75% respectively, so they have a

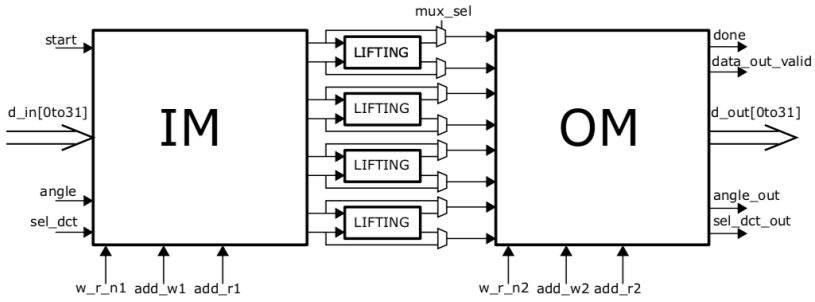

Fig. 2. Steerable block structure

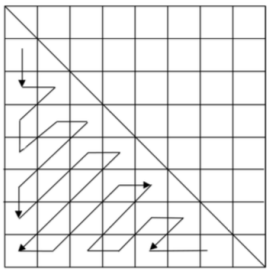

Fig. 3. Zig-zag scanning order

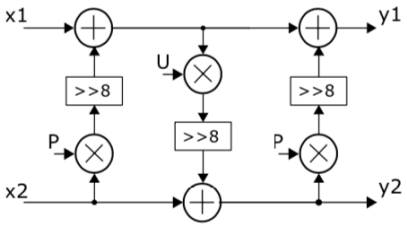

Fig. 4. Lifting-based rotation

parallelism of 16 or 8 data instead of 32, reducing the size of all the memories. In particular the length of both rows and columns of all memories is halved in the SDCT-16 unit, while is four time lower in the SDCT-8 unit with respect to SDCT-32. As a result the area occupation of these units is much lower than the SDCT-32 one. Moreover, just one clock domain has been used for both DCT and steerable block.

3 Results

In order to satisfy the HEVC speed requirements for a video resolution of 7680×4320 and a frame rate of 60 fps, the proposed structure needs a throughput of almost 3 GSample/s. As discussed in Section 2, the folded version presented in [5] has been implemented since this approach guarantees the required throughput. This structure has a processing rate of 16 pixels per cycle, therefore the architecture needs a frequency of at least 187 MHz ($2.99 \times 109/16$ MHz). Clock gating has been enabled for the synthesis, leading to a smaller area and lower power consumption. The technology employed for the synthesis is the UMC 65 nm. The following architectures have been considered and synthesized:

- two-dimensional DCT
- SDCT
- reduced SDCT-16
- reduced SDCT-8

For the SDCT implementation, several clocks have been tested for the steering part, namely 1x, 2x, 4x and 8x. By increasing the Steerable unit frequency it is possible to decrease the parallelism and consequently the number of input/output ports of the buffers.

Cell	1x total area	2x total area	4x total area	8x total area
SDCT	4337744 μm^2	3042226 μm^2	1608759 μm^2	1301522 μm^2
2D-DCT	438866 μm^2	601970 μm^2	455150 μm^2	474167 μm^2
IM	1401523 μm^2	820032 μm^2	495856 μm^2	335932 μm^2
OM	2377837 μm^2	1418162 μm^2	482048 μm^2	319037 μm^2
FIFO	86542 μm^2	110594 μm^2	113008 μm^2	110604 μm^2
ROM	5895 μm^2	22228 μm^2	13227 μm^2	33223 μm^2

Table 1. SDCT area occupation for different clock regimes

It can be noticed that by reducing the data parallelism of the Steerable unit, the size of the input memory (IM) and output memory (OM) decreases considerably, while the size of all the other sub-blocks slightly increases.

Power	Internal	Switching	Total dynamic	Leakage
basic DCT	36.55 mW	17.72 mW	54.47 mW	33 μW
clock gated DCT	21 mW	12.52 mW	33.52 mW	30 μW
basic SDCT	290.47 mW	60.33 mW	350.88 mW	106 μW
clock gated SDCT	88.71 mW	59.85 mW	148.67 mW	94 μW
clock gated SDCT-16	27.86 mW	28.97 mW	56.85 mW	27 μW
clock gated SDCT-8	6.56 mW	7.20 mW	14.17 mW	7 μW

Table 2. Estimated power consumption at 188 MHz

In literature there are no other SDCT hardware architectures, so it is not possible to make comparisons. However, Table 3 presents an overview of the obtained results. As it can be noticed, the area and power results of the SDCT-16 are around 60% smaller than the complete SDCT. On the other hand, the SDCT-8 area is around 75% smaller than the SDCT-16 and 90% smaller than the complete SDCT while the throughputs are reduced respectively by 50% and 75%. Finally, comparing the DCT and the SDCT architecture we can observe that the hardware overhead to support up to N=32 is very large. However, removing the hardware support for the steering part with N=32 (SDCT-16), the area becomes comparable with the one of the DCT. As a consequence, this solution can be of interest to increase the rate-distortion performance [4].

Architecture	DCT	SDCT	SDCT-16	SDCT-8
Technology (nm)	65	65	65	65
Frequency (MHz)	188	188	188	188
Power (mW)	33.52	148.67	56.85	14.17
Throughput	$2.992G$	$2.992G$	$1.496G$	$0.748G$
Area (mm^2)	0.321	1.427	0.444	0.110

Table 3. Overview of the obtained architectures

4 Conclusion

This paper provides an efficient and compact hardware architecture accelerator for the SDCT algorithm to be used in the HEVC algorithm. Many of the design choices explained above present an optimized approach, such as the lifting-based approach, in which the hardware resources are reduced to a minimum. Moreover, the flexibility showed by this architecture makes it appealing for a wide range of applications, being able to work with different coding formats. The proposed SDCT framework is able to cope with 8k UltraHigh Definition (UHD) ($7680 \times 4320 pixels$) with a frame rate of 60 Hz for the 4:2:0 YUV format, which is one of the highest resolution supported by HEVC. The steerable DCT is a viable solution to improve compression efficiency, as reported in [4]. Further work will cover the integration of the proposed accelerator in a complete HEVC framework to validate the performances in a real case scenario.

References

1. Chen, Z., Han, Q., Cham, W.: Low-complexity order-64 integer cosine transform design and its application in hevc. IEEE Transactions on Circuits and Systems for Video Technology 28(9), 2407–2412 (Sep 2018)
2. Daubechies, I., Sweldens, W.: Factoring wavelet transforms into lifting steps. Journal of Fourier Analysis and Applications 4(3), 247–269 (May 1998), https://doi.org/10.1007/BF02476026
3. Fracastoro, G., Fosson, S.M., Magli, E.: Steerable discrete cosine transform. IEEE Transactions on Image Processing 26(1), 303–314 (Jan 2017)
4. Masera, M., Fracastoro, G., Martina, M., Magli, E.: A novel framework for designing directional linear transforms with application to video compression. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1812–1816 (May 2019)
5. Meher, P.K., Park, S.Y., Mohanty, B.K., Lim, K.S., Yeo, C.: Efficient integer dct architectures for hevc. IEEE Transactions on Circuits and Systems for Video Technology 24(1), 168–178 (Jan 2014)
6. Ogata, J., Ichige, K.: Fast intra mode decision method based on outliers of dct coefficients and neighboring block information for h.265/hevc. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5 (May 2018)
7. Oliveira, R.S., Cintra, R.J., Bayer, F.M., da Silveira, T.L.T., Madanayake, A., Leite, A.: Low-complexity 8-point dct approximation based on angle similarity for image and video coding. Multidimensional Systems and Signal Processing 30(3), 1363–1394 (Jul 2019), https://doi.org/10.1007/s11045-018-0601-5
8. Sullivan, G.J., Ohm, J., Han, W., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on Circuits and Systems for Video Technology 22(12), 1649–1668 (Dec 2012)
9. Sun, H., Cheng, Z., Gharehbaghi, A.M., Kimura, S., Fujita, M.: Approximate dct design for video encoding based on novel truncation scheme. IEEE Transactions on Circuits and Systems I: Regular Papers 66(4), 1517–1530 (April 2019)
10. Zeng, B., Fu, J.: Directional discrete cosine transformsa new framework for image coding. IEEE Transactions on Circuits and Systems for Video Technology 18(3), 305–313 (March 2008)