

VLSI implementation of a bi-processor architecture for generic subband coding

P.Desneux

Alcatel Mietec
Rue de la Fusée, 62
B-1140 Bruxelles

pi_desneux@mietec.be

J.D. Legat

Université Catholique de Louvain
Place du Levant, 3
B-1348 Louvain-la-Neuve

legat@dice.ucl.ac.be

Abstract - This paper presents an architecture for the multiresolution coding of pictures. A VLSI implementation has been realized and can achieve a peak performance of about 500 MOPS. The architecture consists in 2 processors whose complementarity enables to avoid any wait cycles during the execution so that the available computation power is continuously used. Moreover, the circuit has a total programmability with respect to the used filters and the picture format; it also has the possibility to take edge effects into account and therefore improve the coding performances. The circuit can be used in the coding as well as in the decoding.

1 Introduction

Many image compression schemes using the subband coding have been proposed for some years [1, 2]. The subband coding is closely related to the multiresolution decomposition of a picture and has many advantages with regard to the DCT-based techniques. Associated to an adapted entropy coder, it can also achieve higher compression ratios. Moreover, the multiresolution transform is close to the human visual system and achieves a high perceptual quality.

The multiresolution has thus very attractive and interesting features, but up to now, it is not part of any standardized coding systems. Many parameters can still be tuned and optimized in order to achieve the best performances in the different application fields. This optimization and research often call for real-time tests that can not be easily performed without dedicated VLSI. This paper proposes such a programmable VLSI architecture.

2 The multiresolution transform

Figure 1-a depicts the subband coding of a picture consisting in successive stages where the image is splitted in four subbands in the 2D frequency space. To achieve higher quality with an identical transmission bit rate, the LL-subband can be recursively splitted in higher order stages. The compression ratio is adjusted with the quantization step. Each stage (Figure 1-b) implements a 2D filter and the proposed architecture uses separable filters with 1D lattice filters (Figure 1-c)[3]. Such a lattice L -tap filter is composed of $N = \frac{L}{2} - 1$ cross-sections characterized by the filter coefficients $\{\rho_0, \dots, \rho_{N-1}\}$, each cross-section performing the following computations : $out_{up} = inp_{up} + \rho * inp_{low}$ and $out_{low} = inp_{low} + \rho * inp_{up}$. The last section of the filter consists in an adder (resp. a subtractor) that outputs the low (resp. high) frequency band. Finally, a scaling multiplier is put on each branch in order to verify the perfect reconstruction property.

The lattice structure is based on a polyphase decomposition of QMF filter banks and is very desirable on a coding viewpoint as it enables the perfect reconstruction property ($\hat{x}_0(n) = x_0(n)$). As far as the VLSI implementation is concerned, the lattice structure has many advantages as well. At first, the

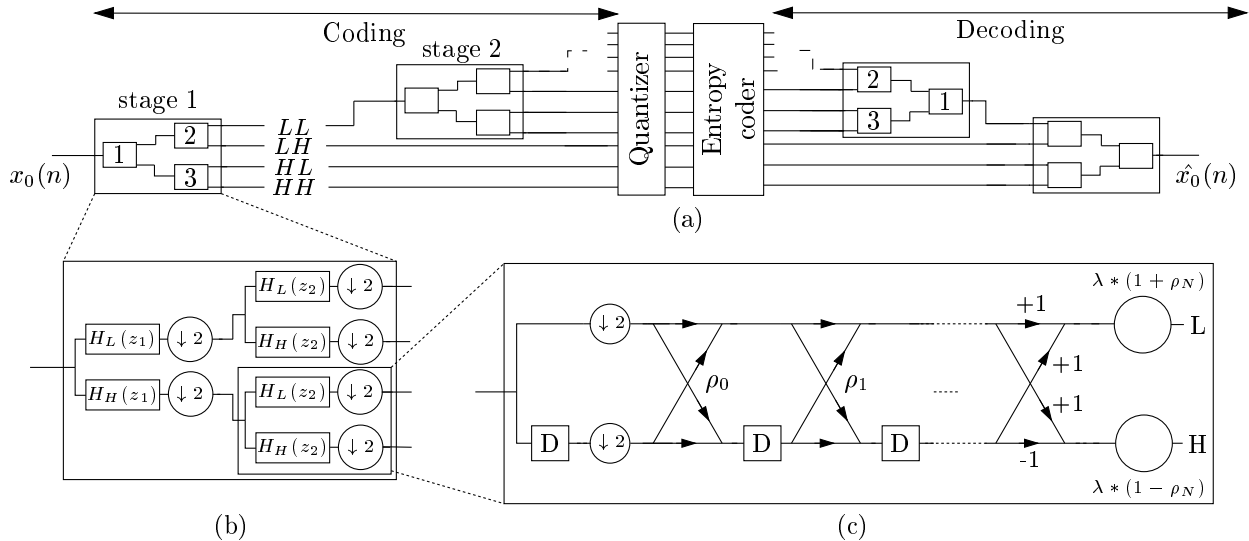


Figure 1: (a): the multiresolution global scheme
 (b): a separable multiresolution stage : $H_L(z)$ =low-pass filter, $H_H(z)$ =high-pass filter, z_1 =vertical direction, z_2 =horizontal direction
 (c): 1D lattice filter of length $L = 2 * (N + 1)$ of the coding part. A similar structure is used in the decoding.

polyphase decomposition and the existing relation between $H_L(z)$ and $H_H(z)$ enable to reduce the required computation power. Compared to a non-polyphase implementation based on direct form I filter structures, the computation power is divided by 4. Secondly, [4] shows that the lattice structure has a high resistance to the roundoff noise caused by finite wordlengths. These wordlengths at the different locations in the multiresolution structure are dimensioned in [4] in order to maximize the precision. Table 1 gives the required computation power for a CCIR601 television signal with a filter length $L = 8$.

	+/- 16 bits	+ 24 bits	* 8b x 16b	* 10b x 16b
1 stage	20.7	62.0	62.0	10.3
3 stages	28.5	85.0	85.0	14.2

Table 1: Computation power (in MOPS)

3 The multiresolution architecture

Figure 2 depicts the proposed architecture which consists in two independently controlled processors and a data memory. The high efficiency of the architecture is based on the tasks distribution between the 2 complementary processors and the adaptation of each processor to its specific tasks. The **transfer processor** (TP) acts as an intelligent DMA coprocessor and provides the **computation processor** (CP) with the data. The CP performs all the computations required in the lattice structures of the different multiresolution stages.

An important architecture feature is its flexibility and programmability. Both processors have their own microprogram RAM which is externally loaded at boot time and a microcontroller managing the program execution. Both microcontrollers are based on a sequencer specifically adapted to the kind of instructions generally used in signal processing. In particular, loop counters enabling up to 4 nested loops levels are hardware implemented. Moreover, the controller of the TP has an interruption controller taking into account 2 different hardware interruptions related to the data inputs or the end of lines. The programmability allows the architecture to be used in several applications (lossless compression for medical pictures archiving, low bit-rate multimedia (like MPEG-4), ...) with different requirements. Another interesting feature is the possibility to use this architecture in picture coding as well as in decoding. Indeed, the structures of an analysis or synthesis stage are very similar and only differ in the order to perform the operations. This can easily be taken into account in the microprograms.

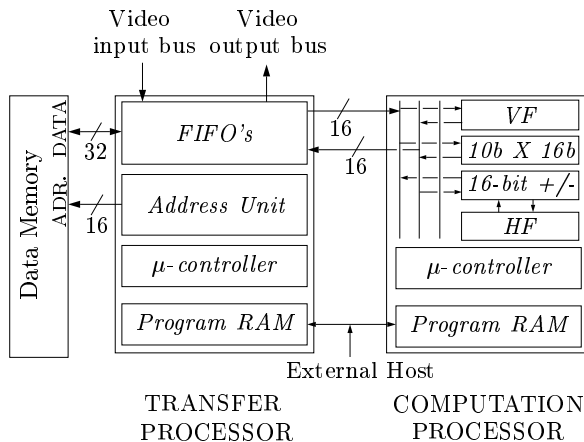


Figure 2 : The multiresolution architecture

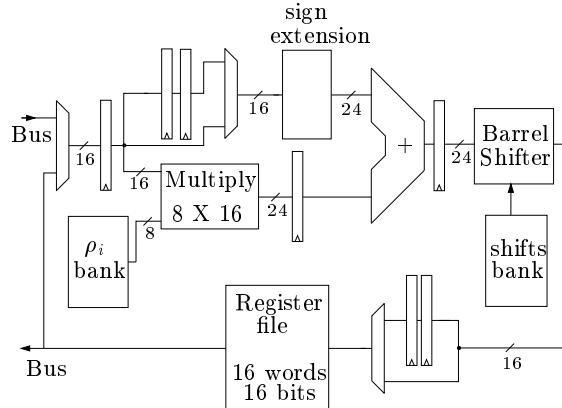


Figure 3 : The VF and HF resources

The desired flexibility explains the different choices made in the architecture design. In particular, the programmability increases the complexity of the data accesses task and therefore requires a dedicated processor with high control capabilities. Programmable parameters such as the filter lengths, the stages number and the picture format influence the data accesses number and order. A CCIR 601 television signal multiresolution transform with typical parameters values ($L=8$, 3 stages) requires a data memory throughput of about 950 Mbits/s. The proposed bi-processor architecture with a dedicated transfer processor is well suited to handle such high bit rates with sufficient flexibility. The data memory is mainly used to store temporary results of the filtering along the vertical direction. A multiplexing scheme between the different picture columns is used and requires line delays instead of single unit delays in the lattice structure (Fig. 1-c). Using hardware line delays could no provide enough flexibility and our architecture is therefore based on a single memory.

The *transfer processor* performs all the tasks related to the data transfers. It has an address computing unit and a FIFO's block. The address unit provides the data memory with the 16-bit address and consists in an adder and a 16-word triple port (1 input, 2 outputs) register file containing the pointers to the different line delays. The FIFO's block consists in two 16-bit x 16 words FIFO's through which data are transferred to and from the computation processor. The FIFO's enable the synchronization between the 2 processors. In order to reduce the memory access rate, the memory wordlength has been fixed to 32 bits. The TP has resources to make the format transforms between the memory words and the CP words.

The high control capabilities of the TP enable to take into account the larger irregularity and complexity of the data accesses in comparison with other algorithms like the DCT or the block matching. Two algorithm specificities explain this complexity: on one part, the edge effects and on the other part, the implementation of multiple multiresolution stages on a single resource. The edge effects consist in the signal extension at the edges and are required in the digital filtering of finite length signals. Several extension methods [5] are possible and can greatly improve the performances especially when the filter lengths are not inconsiderable with regard to the picture size (which can occur in multistage transforms). For a multistage transform, the operations related to the different stages are temporally multiplexed on the resources of the CP that can only implement a single stage. The basic multiplexing block consists in 2 successive lines of a resolution and the scheduling of the different blocks is controlled by the TP. The scheduling scheme is put in the TP microprogram.

The *computation processor* has 4 distinct resources working in parallel. These resources are connected through a dense programmable interconnection network that is based on 3 splittable busses. This connection network is a key point of the CP as it enables many data exchanges between the resources in a single clock cycle. This allows to execute every clock cycle up to 8 operations combined with their related data accesses. The CP achievable computation power is about 500 MOPS.

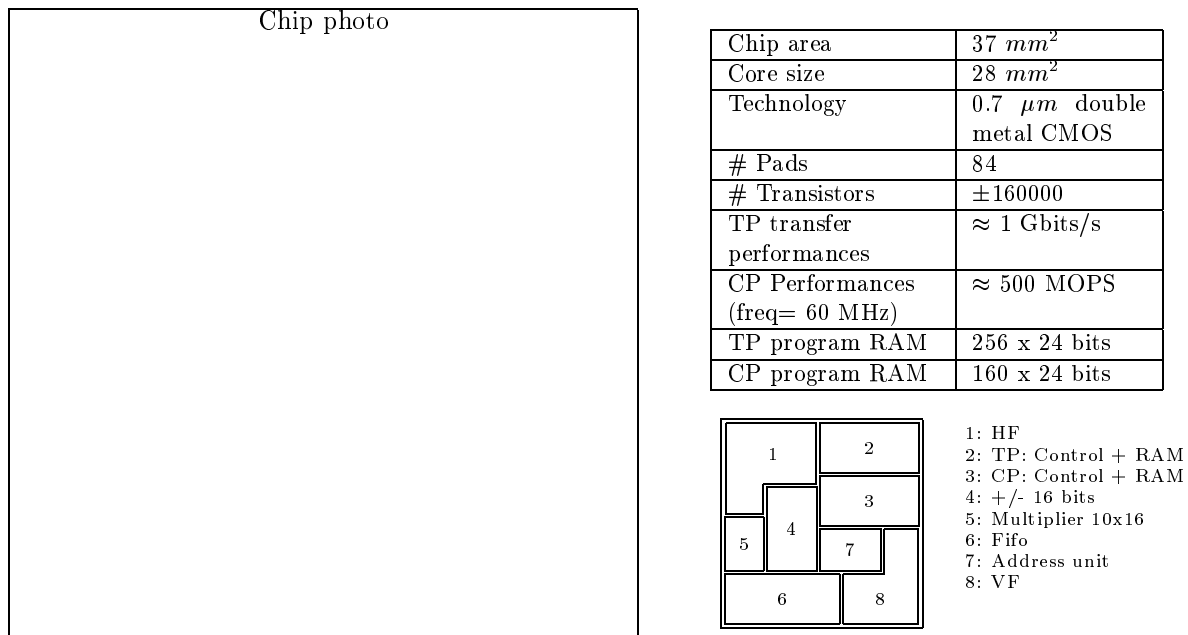
Combined together, the 4 CP resources implement a single multiresolution stage. According to the separability property, the filterings in the vertical (block1 in Fig. 1-a) and horizontal (blocks 2 and 3) directions are made by two distinct resources, respectively VF and HF. The required computation power

is identical for both filters as the outputs of the block 1 are downsampled. However, these resources do not implement the full lattice structure of Figure 1-c but only the ρ cross-sections. The addition-subtraction cell at the end of the 3 blocks is implemented by a single 16-bit adder-subtractor. In order to reduce the multiplications number, the outputs multipliers of the block 1 can be transferred at the outputs of blocks 2 and 3 and combined with the multipliers of these blocks. A single 10X16 multiplier performs all these scaling multiplications.

Figure 3 depicts the VF and HF resources. The pipelined structure allows to execute all the operations of a sub-crossing in 2 successive clock cycles. This corresponds to a computation power of 180 MOPS at 60 MHz. (L-2) cycles are used to implement a L-tap filter successively executing the cross-sections with the different ρ_i stored in a register file. The barrel shifter is used to keep a constant pixel width in the lattice filter according to a scaling method [4] based on programmable power-of-2 shifts.

4 VLSI implementation

A VLSI implementation of the architecture has been realized. For cost reasons, the data memory is implemented in an external static RAM. The following figure shows a microphotograph of the circuit and gives its main properties.



5 Acknowledgments

The authors would like to acknowledge the financial support of the National Fund for Scientific Research in Belgium and of the Walloon Region.

References

- [1] J. WOODS and S. O'NEIL, "Subband coding of images," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 1278–1288, Oct. 1986.
- [2] M. ANTONINI, M. BARLAUD, P. MATHIEU, and I. DAUBECHIES., "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, Apr. 1992.
- [3] P. VAIDYANATHAN, "Multirate digital filters, filter banks, polyphase networks and applications: a tutorial," *Proceedings of the IEEE*, vol. 78, pp. 56–93, Jan. 1990.
- [4] P. DESNEUX, J. MERTES, B. MACQ, and J. LEGAT, "A scaling method for linear-phase lattice filters for multiresolution image coding," in *SPIE Proc. Visual Comm. and Image Proc.*, pp. 1788–1799, SPIE, Sept. 1994. Chicago (USA).
- [5] H. BARNARD, *Image and video coding using a wavelet decomposition*. PhD thesis, TU Delft (The Netherlands), May 94.