Efficient VLSI Architecture for Real-Time Motion Estimation in Advanced Video Coding

Tiago Dias
tdias@sips.inesc-id.pt
DEETC-ISEL / INESC-ID
Lisbon, Portugal

Nuno Roma
nuno.roma@inesc-id.pt
DEI-IST / INESC-ID
Lisbon, Portugal

Leonel Sousa
las@inesc-id.pt
DEEC-IST / INESC-ID
Lisbon, Portugal

ABSTRACT

This paper proposes a new scalable and efficient VLSI architecture for sub-pixel motion estimation. Based on this architecture, a modular and fully configurable motion estimation co-processor is also presented. The efficiency of such processing structure was assessed by embedding this circuit in a half-pixel accurate hierarchical motion estimation system using a two-step search procedure. Experimental results using FPGA devices show that the proposed motion estimation co-processor allows the estimation of motion vectors with half-pixel accuracy in real-time for the 4CIF image format.

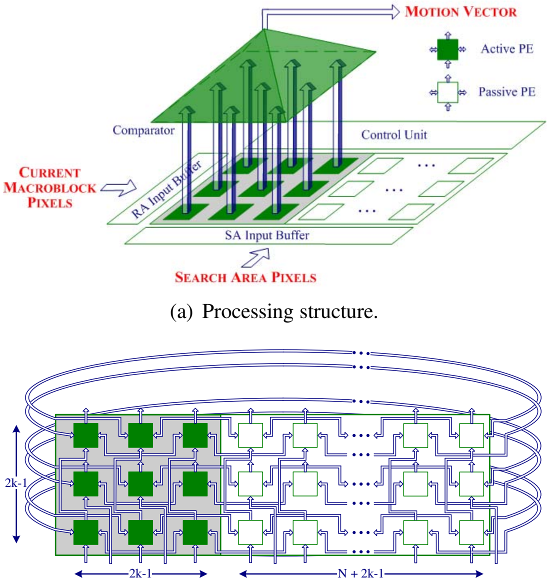
I. INTRODUCTION

A new efficient and scalable VLSI architecture for real-time sub-pixel motion estimation (ME) is proposed. This architecture requires only three sets of data for its operation: the initial coarser motion vector (MV) coordinates, the search area pixels surrounding that point and the current macroblock pixels. The coordinates of the starting point of the sub-pixel ME procedure can be computed in any other hardware or software application. Consequently, this novel architecture can be used to improve the accuracy of the ME process or to estimate local motion based on MVs predicted from previous frames, in both hardware and hybrid software-hardware video coding systems.

II. SINGLE ARRAY ARCHITECTURE FOR ME WITH SUB-PIXEL ACCURACY

In this new type-II structure [1], whose block diagram for the case of half-pixel accuracy is depicted in Fig. 1(a), $(2k-1)^2$ active processing elements (PEs) compute in parallel the cost functions for all candidate blocks, where k is the sub-pixel accuracy (SPA) factor ($k = 2$ for half-pixel, $k = 4$ for quarter-pixel, and so on). In each clock cycle, all active PEs are fed with the same pixel of the macroblock under processing, but with different search area pixels matching the same relative location for different candidate blocks. As a result, in the proposed structure all the cost functions become simultaneously available at the PEs outputs after N^2 processing cycles. These values are then compared against each other in a binary tree comparator in order to find the optimum displacement vector.

To optimize the processing, the proposed architecture has also a set of $(2k-1) \times (N+2k-1)$ *passive PEs*, which are mainly composed by data registers and that are used to store and displace the search area pixels. Furthermore, to guarantee that these search area pixels are displaced over a sampling grid with $\frac{1}{k}$-pixel resolution, all PEs are connected to each other in an interleaved manner in groups of N in the horizontal direction and in groups of k in the vertical direction, being spaced by $k-1$ PEs in both directions, as shown in Fig. 1(b). Moreover, to minimize the hardware requirements of the proposed structure, i.e., to avoid redundant processing elements, the passive PEs located in the right margin of the passive block were also connected to the active PEs in the left margin of the active block, thus creating a cylindrical structure. By adopting this regular interconnection scheme and cylindrical structure, it is possible to use the zig-zag processing scheme proposed in [2] and further optimize the efficiency of the proposed architecture: all PEs are kept busy at any time instant and both the redundant accesses to the frame memory and the need for dummy clock cycles between adjacent rows of search area pixels are avoided.



(a) Processing structure.

(b) PEs interconnections.

Figure 1: Proposed type-II architecture for sub-pixel ME, considering half-pixel accuracy ($k = 2$).

III. SUB-PIXEL MOTION ESTIMATION CO-PROCESSOR

A. *Architecture*

The proposed single array type-II architecture was used as the basis for a modular and fully parameterizable sub-pixel ME co-processor. This circuit was designed using a highly flexible and scalable architecture and makes use of simple and efficient interface protocols to transfer the data between its several units. A pipeline processing scheme was also adopted to interconnect the several modules of the system. As a result, the proposed architecture not only maximizes the data throughput but also provides an easy integration and development of new and different functional blocks.

The proposed circuit is mainly composed by four distinct modules: the ME structure described in Section II.; an interpolation unit, structured in a flexible and modular way so as to being able to implement any generic interpolation algorithm; and two input buffers, that feed the processing array with pixels from the current and candidate macroblocks, respectively. Furthermore, this structure is also fed with the coarser integer-pixel accurate (IPA) MV coordinates, which are used by the search area input buffer to retrieve the subset of pixels required by the interpolation process. These pixels can be retrieved either from the main frame memory or from an intermediate local memory, thus avoiding redundant accesses to the main system memory [3].

B. Implementation and Experimental Results

The proposed sub-pixel ME co-processor was completely described using both behavioural and fully structural parameterizable VHDL. Several setups of these descriptions, considering the typical parameters adopted in videoconferencing applications (half-pixel accuracy and the bilinear interpolation algorithm), were synthesized using the Xilinx Synthesis Tool from ISE 6.1.3i and implemented in a general purpose Virtex XCV3200E-7 FPGA, which has approximately 4 million system gates. The implementation setup presented in this paper considers 8-bit to represent the pixel values and macroblocks with 8 × 8 ($N = 8$) and 16 × 16 pixels ($N = 16$), the set of parameters more frequently adopted by the ITU-T H.26x and ISO MPEG-x video coding standards.

From the experimental results presented in Table 1 it is possible to conclude that the proposed architecture requires very few hardware resources for its implementation. Moreover, considering that in this structure the pixel rate is equal to the clock rate and that the maximum operating frequency values presented in Table 1 impose no clock-rate constraints using nowadays existing technology, it is also possible to conclude that the proposed structure is able to estimate MVs with half-pixel accuracy in real-time for the 4CIF image format. By taking into account the adopted pipeline systolic structure, one can also predict that more accurate MVs can be obtained by merely increasing the number of PEs in the processing array, which will only change the latency of the circuit due to an augment in the number of levels of the comparison unit. Consequently, from the experimental results presented in Table 1 it can be concluded that the proposed sub-pixel ME architecture is able to estimate MVs with eighth-pixel accuracy in real-time for any CIF image format, with a minor increase of its latency.

Table 1: Percentage of CLB slices, LUTs and maximum operating frequency of the proposed ME system with half-pixel accuracy in a Virtex XCV3200E-7 FPGA.

	$\mathbf{N = 8}$			$\mathbf{N = 16}$		
Architecture Type	**CLBs**	**LUTs**	**F[MHz]**	**CLBs**	**LUTs**	**F[MHz]**
SPA	3.8%	3.2%	141.4	5.2%	4.5%	138.8
IPA + DR	21.5%	10.4%	129.4	83.9%	38.4%	124.1
IPA + DR + SPA	27.9%	14.6%	129.4	95.1%	45.6%	124.1

IV. VALIDATION ON PRACTICAL IMPLEMENTATION

To validate the functionality of the proposed sub-pixel ME architecture, a complete ME system that estimates MVs with half-pixel accuracy using a two-step search algorithm was developed and implemented. The system is composed by two distinct modules: one that estimates IPA MVs and a second one that refines these MVs coordinates into half-pixel resolution. The IPA ME module is based on an efficient class of bi-dimensional processing structures for FSBM ME recently proposed in the ME literature [2]. The SPA ME module consists of an implementation of the sub-pixel ME processor introduced in Section II., considering half-pixel accuracy and the bilinear interpolation algorithm. An important feature of this ME system is that it minimizes the memory bandwidth requirements of the main frame memory, by reusing and transferring the data between its modules through a dedicated data reuse unit (DR) [3].

Table 1 presents the experimental results obtained for several different implementations of the proposed multi-level ME processor with half-pixel accuracy, using the above mentioned setup parameters. From these results it is possible to conclude that the maximum operating frequency of the proposed ME system is not constrained by the sub-pixel ME processor. Consequently, this ME system is able to estimate MVs with half-pixel accuracy for the 4CIF image format up to a rate of 30 fps.

V. CONCLUSION

A new VLSI type-II architecture for sub-pixel ME in video sequences, highly suitable for both hardware and hybrid software-hardware ME systems, was proposed in this paper. Based on the proposed type-II architecture, an efficient modular and configurable sub-pixel ME co-processor was also presented in this paper. Experimental results for FPGA implementations show that using such processor it is possible to estimate MVs with half-pixel accuracy up to a rate of 30 fps for high-quality video (4CIF format).

REFERENCES

1. P. Pirsch, N. Demassieux, and W. Gehrke, “VLSI architectures for video compression - a survey,” *Proc. of the IEEE*, vol. 83, no. 2, pp. 220–246, Feb. 1995.
2. N. Roma and L. Sousa, “Efficient and configurable full-search block-matching processors,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1160–1167, Dec. 2002.
3. T. Dias, N. Roma, and L. Sousa, “Efficient motion vector refinement architecture for sub-pixel motion estimation systems,” in *Proc. IEEE Workshop on Signal Processing Systems (SIPS’05)*, Athens - Greece, Nov. 2005.