Algorithm and Architecture of Prediction Core in Stereo Video Hybrid Coding System

Li-Fu Ding, Shao-Yi Chien, and Liang-Gee Chen
DSP/IC Design Lab, Graduate Institute of Electronics Engineering and
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
Email: {lifu,shaoyi,lgchen}@video.ee.ntu.edu.tw

***Abstract*— 3D video will become noticeable video technology in the next generation. In this paper, a stereo video coding system is proposed from algorithm level to hardware architecture level. We propose a novel stereo video coding system by exploiting joint block compensation scheme to achieve high coding efficiency. It is also suitable for hardware implementation. Due to more than twice computational complexity relative to mono video coding systems, a new hardware architecture based on hierarchical search block matching algorithm (HSBMA) with some modification is proposed. With special data flow, no bubble cycles exist during block matching process. Proposed architecture also adopts near overlapped candidates reuse scheme (NOCRS) to save heavy burden of data access. Besides, by the proposed new scheduling, both on-chip memory requirement and off-chip memory bandwidth can be reduced. A prototype chip can achieve real-time requirement under the operating frequency of 81 MHz for 30 D1 frames per second (fps) in left and right channel simultaneously, with ME/DE search range of [-64, +63] in horizontal direction and [-32, +31]/[-16, +15] in vertical direction. Compared with the hardware requirement for implementation of full search block matching algorithm (FSBMA), only 11.5% on-chip SRAM and 1/30 amount of PEs are needed. It shows that the hardware cost is quite small.**

I. Introduction

Stereo video can make users have 3D scene perception by showing two frames to different eyes simultaneously. With the technologies of 3D-TV getting more and more mature [4], stereo and multi-view video coding draw more and more attention. In recent years, MPEG 3D auido/video (3DAV) Group has worked toward the standardization for multi-view video coding [5], which also advances the stereoscopic video applications. Although stereo video is attractive, the amount of video data and the computational complexity is doubled. A good coding system is required to solve the problem of huge data with limited bandwidth. Besides, in a mono video coding system, the prediction in temporal domain, motion estimation (ME), requires the most computational complexity [1]. By comparison, computational load of prediction is heavier in stereo video coding systems due to additional ME and disparity estimation (DE), which is the prediction in spatial domain, in the additional channel. For real-time applications, a hardware accelerator solution is urgently required due to the heavy computational complexity.

Under real-time constraint, it is preferred that ME and DE unit, which is called "prediction core" in the system, should be realized by a hardware accelerator due to its heavy computational load. Among all the block matching algorithms, full search block matching algorithm (FSBMA) is the most popular [3]. Many FSBMA architectures were developed based on the regular data flow [7]. However, for D1 30 fps stereo video contents, hardware processing parallelism needed is more than 4096 (that is, 4096 absolute difference operations execute simultaneously) to achieve real-time requirement. The hardware cost is quite large. Hierarchical search block matching algorithm (HSBMA) has been regarded as powerful computational configuration in BMA [3]. It can effectively reduce not only computational load but also on-chip memory size. However, due to its irregular data access in fine levels, data of search windows (SWs) are hard to be reused effectively. It causes much more data access load than FSBMA.

In this paper, an algorithm and hardware architecture for stereo video coding system with hybrid coding scheme are proposed. The algorithm is designed for hardware implementation. It is modified from our prior successful algorithm [2]. We exploit joint block compensation to achieve good coding efficiency of the stereo video system. HSBMA is modified for better video quality, and it is applied in the prediction core design of the system. Then, a new hardware architecture based on the proposed hardware-oriented algorithm is proposed. NOCRS and new scheduling are adopted to reduce data access load and on-chip memory requirement. Besides, the proposed data flow of block matching process (BMP) in three levels of HSBMA is adopted for eliminating bubble cycles.

The hardware-oriented algorithm for stereo video coding system is first described in the next section. Then the hardware architecture and implementation results are shown in Section III and IV, respectively. Finally, in Section V gives the conclusion.

II. Hardware-Oriented Algorithm for Stereo Video Coding System

For the purpose of compatibility, the coding system adopts a base-layer-enhancement-layer scheme, as shown in Fig. 1. The left view is set as the base layer, and the right view is set as the enhancement layer. The base layer is encoded with MPEG-4 video encoder. The implementation of the proposed stereo video coding system is based on two hardware-oriented concepts. First, in order to greatly reduced the area of processing elements (PEs) and on-chip memory, we adopt HSBMA to perform ME and DE. Second, in the compensation step, a block is not only compensated by the

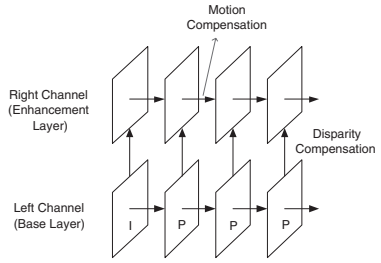

Fig. 1. Base-layer-enhancement-layer scheme of the proposed system.

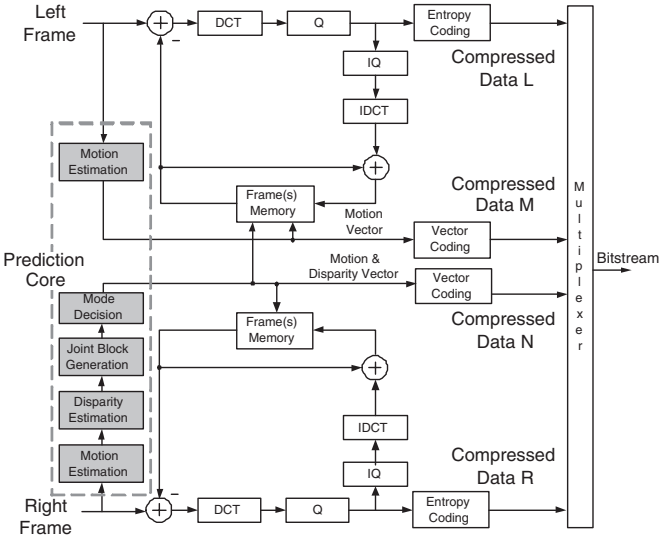

Fig. 2. Block diagram of the proposed stereo video encoder.

block of left or right reference frames, but also the combination of them for different types of content in the current block. Based on these concepts, in this section, the encoding flow is introduced first. Next, the details of modified hierarchical ME/DE algorithm and joint block compensation are shown in the rest subsections.

A. Encoding Flow

The block diagram of the proposed stereo video encoder is shown in Fig. 2. The proposed prediction core mainly includes ME, DE, joint block generation, and mode decision, which is the most computation-consuming part in the system. The main differences between the encoding flows of the left channel and the right channel are DE, joint block generation, and mode decision, which are introduced later. After encoding, the compressed data of the left channel, M and L, and the compressed data of the right channel, which is of a small amount, N and R, are transmitted.

B. Hierarchical ME/DE Algorithm with NOCRS

ME is a key unit in general video coding systems. In the proposed stereo video hybrid coding system, additional ME and DE are required when encoding frames of the right channel. It increases the design challenge in large on-chip memory bandwidth and computational load. Thus HSBMA is adopted and

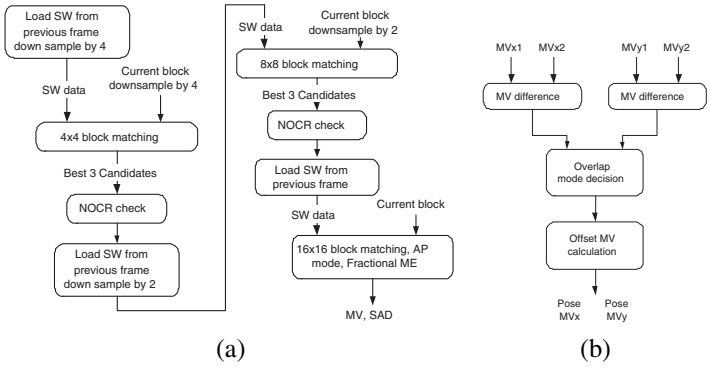

Fig. 3. (a)Flow of proposed hierarchical ME/DE algorithm with NOCRS. (b)Flow of near overlapped candidates reuse scheme (NOCRS).

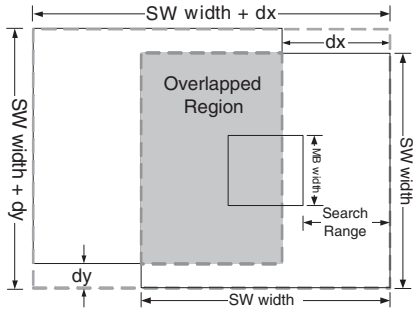

Fig. 4. The union of overlapped search windows.

improved. Figure 3 (a) shows the flow chart of the proposed hierarchical ME/DE algorithm. Once a reconstructed frame is generated in the reconstruction loop in the encoder, it will be passed through a low-pass filter. Then the reconstructed frames down-sampled by two and by four are generated and stored in the off-chip frame buffer. The proposed hierarchical ME/DE algorithm performs three-level BMP. When ME/DE starts, a SW is loaded from the previous frame down-sampled by four to perform 4×4 (level-2) BMP. The best three candidates are chosen for the refinement process in the next level. However, different from FSBMA, SW data cannot be reused effectively in the refinement levels: 8×8 (level-1) and 16×16 (level-0) BMP. It will cause serious overhead on bus bandwidth. However, our experimental analysis shows that MVs of the best three candidates are usually very close. It means that the SWs of them in the next refinement level are also close. Thus, to reduce memory bandwidth requirement, we propose "near overlapped candidates reuse scheme (NOCRS)." Figure 3 (b) shows the flow of NOCRS, after finding the best three candidates during level-2 and level-1 BMPs, three MVs will be checked as follows,

$$MV_{diff}(N) = |MV_1(\mathbf{B}) - MV_2(\mathbf{B})|, \tag{1}$$

$$Overlap(N) = \begin{cases} true, & MV_{diff}(N) < Threshold \\ false, & \text{otherwise.} \end{cases} \tag{2}$$

It is realized by a simple hardware unit which is introduced later. If the overlapped condition is satisfied, instead of being

TABLE I

SYSTEM BANDWIDTH REQUIREMENT OF THREE BMA.

Data loaded from off-chip frame buffer	FSBMA	HSBMA without NOCRS	HSBMA with NOCRS
Current frame	9.9	9.9	9.9
SW for 4x4 BMP	0	3.4	3.4
SW for 8x8 BMP	0	46.4	29.5
SW for 16x16 BMP	55	46.4	31
Reconstructed frame	9.9	9.9	9.9
DS reconstructed frame	0	7.4	7.4
Total (Mega-Bytes / sec)	74.8	123.4	91.1

TABLE II

COMPARISON BETWEEN PROPOSED ALGORITHM AND FSBMA.

Criterion	FSBMA	HSBMA with NOCRS
On-chip memory	180k bits	20.75k bits
Nomalized search points / MB	8192	100 - 234
Parallelism of PEs for real-time	>4096	128
Average quality drop	0	<0.2 dB

loaded separately, a union of these SWs is loaded only once from the off-chip frame buffer, as shown in Fig. 4. NOCRS is a simple but effective scheme. It not only reduces on-chip memory bandwidth successfully but also avoids unnecessary computation on duplicated search candidates in two separately SWs. Table I shows the system bandwidth requirement of three ME/DE algorithms. 35.5% system bandwidth can be saved after NOCRS is applied. Although FSBMA requires less system bandwidth by regular data reuse scheme, for example, level-C data reuse scheme [7], Table II shows the proposed HSBMA with NOCRS has much less on-chip-memory requirement and computational load. At the same time, the proposed algorithm can still maintains good objective and subjective quality, as shown in Fig. 5.

C. Joint Block Compensation Scheme

In ME and DE steps of the right channel, the current block has two reference frames, as shown in Fig. 6. Note that the search range of the left reference frame for DE is not a square because cameras are parallel-structured, so the candidate blocks are only on a belt of region [8]. There are three types of compensated blocks for right channel in the proposed stereo video encoder. 1) Motion-compensated block

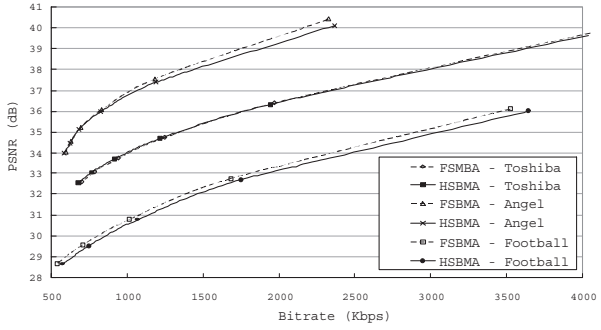

Fig. 5. Comparison of rate-distortion between proposed HSBMA and FSBMA.

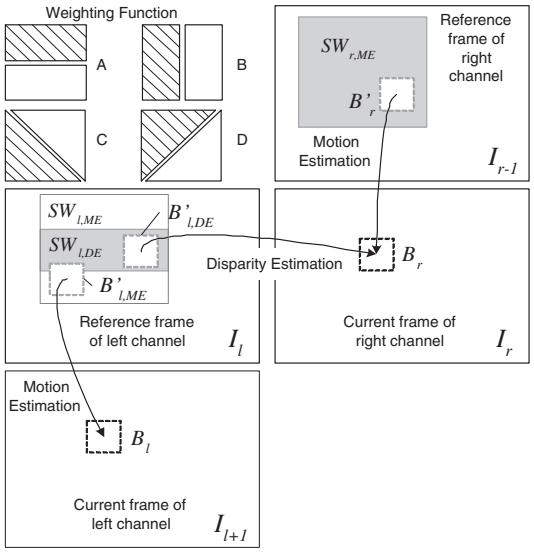

Fig. 6. The illustration of the prediction directions and the search range of two reference frames.

is illustrated as B'_r in Fig. 6: it often occurs in the background due to its zero or slow motion. Occlusions between left and right frames will also be compensated by this type of blocks. 2) Disparity-compensated block is illustrated as $B'_{l,DE}$ in Fig. 6: it often occurs in the moving objects because of their deformation during motion. In this case, disparity-compensated blocks usually have better prediction capability. 3) Joint block: it often occurs in the block which contains object boundary in it because different objects may be suitably predicted by different types of blocks.

According to the criterion of sum of absolute difference (SAD), the best type of compensated block is selected. For each macro block of the current frame, the distortion of the three types of blocks are computed as follows,

$$D_{motion} = \min\{|I_r(B_r) - I_{r-1}(B'_r)| \,|_{B'_r \in SW_{r,ME}(B_r)}\}, \tag{3}$$

$$D_{disaprity} = \min\{|I_r(B_r) - I_l(B'_{l,DE})| \,|_{B'_{l,DE} \in SW_{l,DE}(B_r)}\}, \tag{4}$$

$$D_{j_n} = \sum |\, I_r(B_r) - [W_n \cdot I_l(B'_{l,DE}) + W'_n \cdot I_{r-1}(B'_r)]\,| \quad W_n + W'_n = I, \tag{5}$$

where D_{motion} and $D_{disparity}$ are the minimum SAD values of motion- and disparity-compensated blocks, respectively. $I_r(B_r)$ is the current block in the right channel, $I_{r-1}(B'_r)$ is the reference block in the right channel, and $I_l(B_{l,DE})$ is the reference block in the left channel. $SW_r(B_r)$ and $SW_{l,DE}(B_r)$ are the search windows in right and left reference frames of block B_r, respectively. DE and ME result in the best matching blocks, $I_l(B'_{l,DE})$ and $I_{r-1}(B'_r)$, in the reference frames in the left and right channel. Then the proposed joint block is composed of the weighted sum of

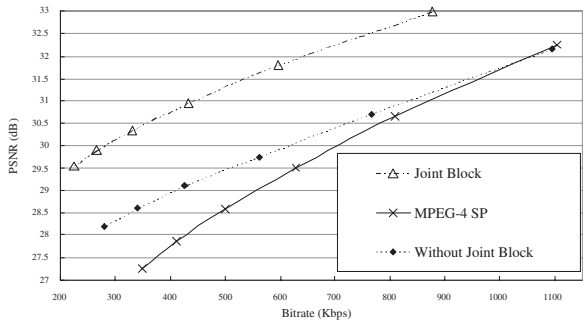

Fig. 7. Rate-distortion curve of sequence "Race2."

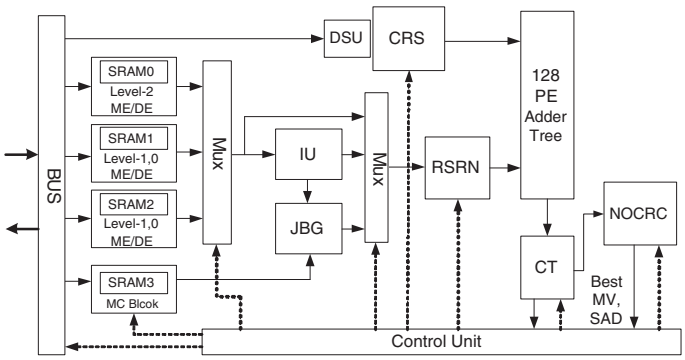

Fig. 8. Architecture of the prediction core chip for stereo video hybrid coding system.

the two blocks, $I_l(B'_{l,DE})$ and $I_{r-1}(B'_r)$. W_n and W'_n are complementary weighting functions that describe weighting parameters. Some example patterns such as A to D, as shown in Fig. 6. In (5), the SAD value D_{j_n} is derived. Finally, the mode decision is described as follows,

$$Mode = \arg\min_{mode}\{D_{motion}, D_{disparity}, D_{j_1}, ..., D_{j_n}\}. \quad (6)$$

As shown in Fig. 7, PSNR degradation is $2-3$ dB without joint block compensation. The proposed scheme greatly improves the coding efficiency of stereo video systems.

III. Hardware Architecture of Prediction Core

The hardware architecture designed for the hardware-oriented algorithm is shown in Fig. 8. There are eight main units: control unit, reference shift register network (RSRN), current register set (CRS), 128-PE adder tree, compare tree (CT), NOCR checker (NOCRC), interpolation unit (IU), and joint block generator (JBG). RSRN is composed of a reconfigurable shift register array. After data loading of search window is finished, RSRN starts to fetch data from on-chip memory. Meanwhile, PE-adder tree accumulates the distortion (absolute difference). Except for several cycles in the beginning for data preparing, SADs of eight candidate blocks in level-2 BMP, two candidate blocks in level-1 BMP, or half candidate blocks in level-0 BMP can be derived in every cycle. Then compare tree can compare these SADs in one cycle. The best three candidate MVs are chosen for refinement process. NOCRC checks the degree of overlapping and outputs the post-processed MVs to address generator (AG) in the control unit, which decides when BMP of the next level begins. IU generates sub-pixels in half pixel refinement process. JBG generates joint block for mode decision for improving coding efficiency of stereo video. The detail architectures of RSRN, JBG, and NOCRC are described in the following three subsections. Furthermore, data reuse scheme and memory organization are also shown. Besides, a new scheduling is proposed to reduce the demand of on-chip memory and bandwidth for data access.

A. Reconfigurable Shift Register Array and Its Data Flow

To achieve the design goal of hierarchical block-matching operation with only one hardware resource, RSRN is composed of a set of reconfigurable shift register array, which

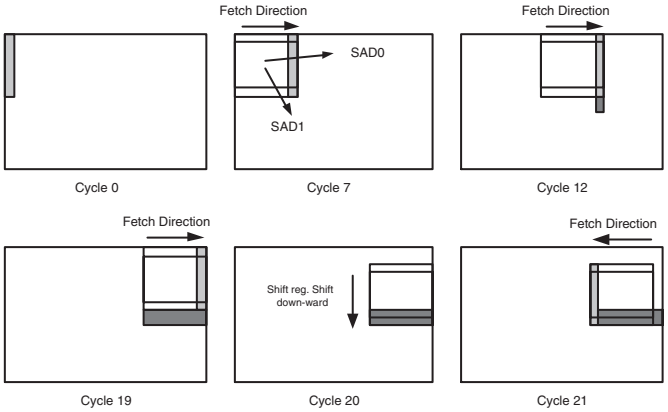

Fig. 9. Data flow of level-1 BMP.

consists of 128 8-bits registers. The outputs of these registers are connected to 128 processing elements (PEs) in the adder tree, which can calculate sum of absolute differences of 128 pixels and accumulate them in one cycle. It has high reconfigurability and can change the connection configuration by the control unit. One column of pixels are fetched from on-chip memory to RSRN every cycle. Because of its reconfigurable feature, data in RSRN can shift left-ward, down-ward, and right-ward, there are no bubble cycles when the search position is changed in the vertical position.

An example of detail data flow is shown in Fig. 9, which is the data flow of level-1 BMP. The search range is [-6, +6] in our design, and the search window is 20×20. When BMP starts, 9×1 search window pixels are inputted each cycle. At cycle 0, the left 9×1 pixels on the column 0 in the search window are inputted. Then the left 9×1 pixels on column 1 are inputted at cycle 1, and go on. At cycle 7, all the candidate block data of search position (-6, -6) and (-6, -5) are stored in RSRN. Thus, the first two 8×8 SADs can be generated at cycle 7. Then, two 8×8 SADs are generated in each cycle. At cycle 12, two additional pixels must be stored in additional registers. At cycle 19, these additional sixteen pixels are ready for inputting to RSRN. RSRN shifts down-ward, and two 8×8 SADs of search position (6, -4) and (6, -3) are generated without any bubble cycles. At cycle 21, RSRN changes the

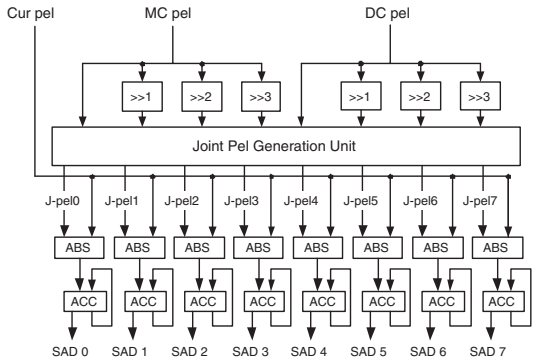

Fig. 10. Architecture of one PE of the joint block generator.

connection configuration again and shifts left-ward. In this way, all the bubble cycles can be avoided besides the beginning several cycles, so the utilization is near 100%. Level-2 and level-0 BMP have the similar flow with level-1 BMP.

B. NOCR Checker

After SADs are generated from 128-PE adder tree, compare tree compares eight 4×4 SADs in level-2 BMP or two 8×8 SADs in level-1 BMP. Both in level-2 and level-1, the best three MVs are chosen, and then inputted to NOCR checker. MV differences are calculated mutually, and then the overlapping condition is decided. For example, if three outputs of Threshold units are all logic 1, it means SW of next level should be loaded only once rather than three times. This architecture is simple, but it can effectively reduce over 30% unnecessary data access from off-chip. Furthermore, it also reduces unnecessary computation and saves processing cycles.

C. Joint Block Generator

When ME of the right channel is finished, the best candidate block must be hold for joint block generation step. As motion compensation process, the best candidate block is loaded into on-chip SRAM "MC block," as shown in Fig 8. After DE of the right channel is also finished, mode decision for joint block starts. Figure 10 shows one of the sixteen processing elements in JBG. Only adders are used to generate weighted sum in joint pel generation unit. In our chip design, there are sixteen PEs to generation eight kinds of joint blocks at the same time. After 16 cycles, eight 16×16 SADs of joint blocks are generated. Then they are outputted to the compare tree to choose the best SAD, and the best mode is derived as well.

D. Data Reuse Scheme

Because of the regular data access of level-2 BMP, level-C reuse scheme is applied for the search window loading in level-2 BMP. The disadvantage of conventional HSBMA [6] is that the search window required for refinement level (level-1, level-0) can not be reused because of its irregular flow. It increases the data access burdens. However, the proposed NOCRS effectively solves this problem. In other words, data reuse scheme are also applied in level-1 and level-0 BMP to save bus bandwidth.

E. Proposed Scheduling for Stereo Video Coding System

The proposed Scheduling is modified from our prior stereo video system [2] for hardware implementation consideration. Figure 11 (a) shows the original frame-based scheduling of the prediction core. ME/DE of right channel cannot start until MVs of all the blocks of a frame in the left channel are derived. However, from Fig. 6 we found that the search window required for DE in the right channel ($SW_{l,DE}(B_r)$, gray rectangular region) is enclosed by the search window required for ME in the left channel ($SW_{l,ME}(B_l)$, white square region). It means in the original scheduling, region $SW_{l,DE}(B_l)$ is loaded twice from the off-chip frame buffer. It wastes bus bandwidth. This problem can be solved after the new scheduling is applied. At the beginning of DE of B_r in the right channel, $SW_{l,ME}(B_l)$ instead of $SW_{l,DE}(B_r)$ is loaded from off-chip. After DE and joint block mode decision are done, ME of B_l in the left channel starts. No loading process is needed before ME. On the other hand, on-chip memory required for $SW_{l,DE}(B_r)$ can be shared with that for $SW_{l,ME}(B_l)$. Therefore, the proposed scheduling reduces both the requirements of off-chip memory bandwidth and on-chip memory.

F. Memory Organization

From Fig. 8, it shows that on-chip SRAM "$SRAM0$" is allocated for level-2 BMP in the left and right channel. With level-C data reuse scheme, data in this SRAM can not be erased after ME/DE of the block is finished. "$SRAM1$" and "$SRAM2$" are allocated for the refinement process (level-1 and level-0 BMP). Because there may be one to three search windows needed for block matching, when $SRAM1$ is fetched pixels by RSRN to perform block matching, $SRAM2$ is loaded another SW data from the off-chip frame buffer, and vice versa. In this way, wasted cycles for loading search window data from off-chip can be greatly reduced. "$SRAM3$" is allocated for holding the best candidate block after ME in the right channel, which is used for joint block generation. In our chip design, only 20.75k bits on-chip SRAM are required, which is only 11.5% on-chip SRAM requirement compared with FSBMA.

IV. Implementation

The design goal of our chip implementation is listed as follows: 720×480 frame size and 30 frames both in the left and right channels. In ME case, the search range is [-64, +63] in the horizontal direction and and [-32, +31] in the vertical direction. While in DE case, the search range is [-64, +63] in the horizontal direction and [-16, +15] in the vertical direction. The chip is designed in cell-based design flow with Artisan 0.18um 1P6M standard cell library and Artisan RAM compiler. The prediction core chip is currently under fabrication by TSMC. The chip layout is shown in Fig. 12. There are three groups of on-chip single-port SRAM on the chip. The functionality of these SRAMs are described in the previous section. The technology is TSMC 0.18um 1P6M, and the chip size is 1.71×1.71mm^2. It shows that the chip size

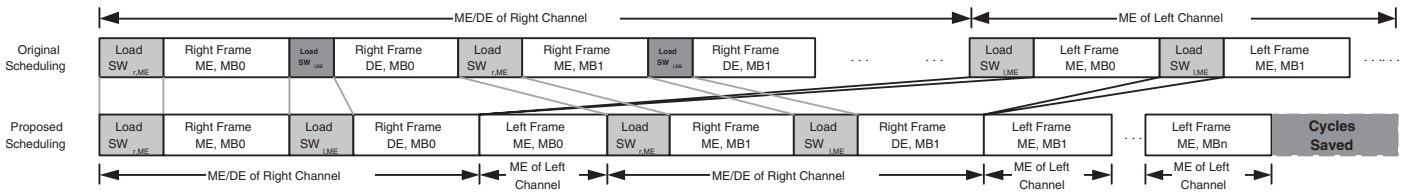

Fig. 11. Original and proposed scheduling of the stereo video system.

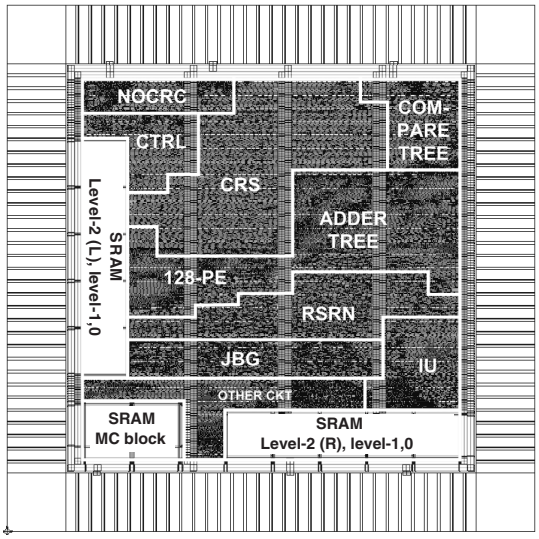

Fig. 12. Chip layout of the proposed prediction engine architecture.

TABLE III
CHIP SPECIFICATIONS.

Technology	TSMC 1P6M 0.18um
Chip size	2.13mm×2.13mm
Package	128 CQFP
On-chip memory	21,248 bits
Logic gate count	137,838
Maximum frequency	100 MHz
Power supply	1.8V
Power consumption	95.85 mW @ 100 MHz
Search range	ME: horizontal [-64, +63], vertical [-32, +31] DE: horizontal [-64, +63], vertical [-16, +15]
Processing capability	30 D1(720x480) frames/sec in both channels including 2 ME and 1 DE operations

is quite small under the high specification and can be easily integrated into a single chip stereo or mono video system. The detailed features of the chip is shown in Table III.

Simulation results show that this chip can achieve real-time requirement for D1 (720×480) stereo video system at 81MHz in general video case. Due to the irregular (cycle-variant) property of hierarchical ME/DE block matching algorithm, the maximum working frequency is designed at 100MHz to handle the worst case in real time. Note that the chip can perform two ME operations of the left and right channel and one DE operations of the right channel in 1/30 second. Compared with the hardware requirement for implementation of FSBMA, only 11.5% on-chip SRAM and one-thirty amount of PEs are needed in this chip. It is quite area-efficient. Algorithm analysis also shows that it maintains good video quality.

V. CONCLUSION

A hardware-oriented stereo video prediction algorithm and its associated hardware architecture is proposed in this paper. Compared with FSBMA, the proposed HSBMA greatly reduces hardware resource requirement (on-chip SRAMs and PEs), while it still maintains good video quality. With NOCRS, the problem of critical memory bandwidth requirement can be solved by checking the overlap degree of three SWs in the refinement level. Joint block compensation utilizes the weighted sum of motion- and disparity-compensated blocks. It can improve the coding efficiency of the stereo video system, and it is easy for hardware implementation. The hardware architecture is design for the proposed algorithm with a set of reconfigurable shift register array, which can be configured for all the scan directions of three block matching level. Thus, no bubbles exist during three level BMPs. Moreover, the proposed scheduling not only reduces cycles for loading data from off-chip frame buffer but also eliminates on-chip SRAM for level-2 of DE. A prototype chip is currently under fabrication with 0.18um 1P6M technology by TSMC. It shows the chip size is small and can be easily integrated into stereo video hybrid coding systems and existing mono-video coding systems.

REFERENCES

[1] H.-C. Chang, L.-G. Chen, M.-Y. Hsu, and Y.-C. Chang, "Performance analysis and architecture evaluation of mpeg-4 video codec system," in *Proceedings of 2000 IEEE International Symposium on Circuits and Systems (ISCAS 2000)*, 2000.

[2] L.-F. Ding, S.-Y. Chien, Y.-W. Huang, Y.-L. Chang, and L.-G. Chen, "Stereo video coding system with hybrid coding based on joint prediction scheme," in *Proceedings of 2005 IEEE International Symposium on Circuits and Systems (ISCAS 2005)*, 2005.

[3] Y.-W. Huang, "Algorithm and architecture design for motion estimation, h.264/avc standard, and intelligent video signal processing," Ph.D. dissertation, Nation Taiwain University, Taipei, Dec. 2004.

[4] F. Isgrò, E. Trucco, P. Kauff, and O. Schreer, "Three-dimensional image processing in the future of immersive media," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 388–303, Mar. 2003.

[5] *Requirements on multi-view video coding*, MPEG-4 Std. ISO/IEC JTC1/SC29/WG11 N6501, 2004.

[6] B.-C. Song and K.-W. Chun, "Multi-resolution block matching algorithm and its vlsi architecture for fast motion estimation in an mpeg-2 video encoder," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 12, no. 9, pp. 1119–1137, 2004.

[7] J. C. Tuan, T. S. Chang, and C. W. Jen, "On the data reuse and memory bandwidth analysis for full-search block-matching vlsi architecture," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 12, no. 1, pp. 61–72, Jan. 2002.

[8] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communication*. Prentice Hall, 2001.