Stereo Video Coding System with Hybrid Coding Based on Joint Prediction Scheme

Li-Fu Ding, Shao-Yi Chien, Yu-Wen Huang, Yu-Lin Chang, 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, yuwen, ylchang, lgchen}@video.ee.ntu.edu.tw

Abstract— **Stereo video systems require double bandwidth and more than twice computational complexity relative to mono-video systems. Thus, An efficient coding scheme is necessary for transmitting stereo video. We propose a novel stereo video coding system by exploiting joint prediction scheme which combines three prediction schemes to achieve high coding efficiency and low computational complexity. Joint block compensation improves the visual quality. Motion vector prediction and mode pre-decision utilize the features of stereo video to reduce the computational complexity with up to 8–9 times acceleration. Experiments show that the proposed joint prediction scheme is 2dB better than MPEG-4 TS and 3dB better than MPEG-4 SP.**

I. Introduction

Stereo video can make users have 3D scene perception by showing two frames to each eye simultaneously. With the technologies of 3D-TV getting more and more mature [5], 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 [2], which also makes advancement of 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, motion estimation (ME) requires the most computational complexity [3]. By comparison, computational loading is heavier in stereo video coding systems due to additional ME and disparity estimation (DE). Thus, an efficient prediction scheme is required to overcome these problems. Finally, it is preferred that the proposed video encoding system is easily integrated by existing video standards.

In the past years, some stereo video coding systems are proposed. Stereo video coding can be supported by temporal scalability tools of existing standards, such as MPEG-2 multiview profile [1]. However, it cannot achieve good coding efficiency. I3D [7] is a famous approach, where the texture information is contained in a synthetic view, and the depth information is contained in a disparity map. It has good coding efficiency and compatibility with existing standard. However, additional operations for extracting disparity maps and synthesizing stereo views are required in the encoder and the decoder, respectively, which are not building blocks of conventional video coding systems. A mesh-based and block-based hybrid approach is proposed [9]. It has good compatibility and achieves acceptable coding efficiency. It needs additional preprocessing for segmentation to prevent matching failure around object boundary. In addition, the computational complexity is very high.

In this paper, we propose a new stereo video coding system with joint prediction scheme for general stereo video applications. The joint prediction scheme contains three coding tools. First, The joint block compensation is employed for better subjective and objective quality. Second, a new motion vector prediction algorithm is proposed according to the features of stereo video. Then, a mode pre-decision scheme is adopted to reduce computational complexity. The rest sections are organized as follows. Section II describes the proposed stereo video coding system. The experimental results are shown in Section III. Finally, Section IV gives the conclusion.

II. Proposed 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 proposed stereo video coding system is based on three concepts. First, in the compensation step, a block is not only compensated by the block of left or right reference frames, but also the combination of them due to different types of content in the current block. Second, in order to reduce the computational complexity of ME, the properties of stereo video, which is introduced later, are considered. It is adopted for accurate motion vector prediction. Third, the computational complexity of DE can be reduced for the similar reason with mode pre-decision scheme. Based on these three concepts, in this section, the encoding flow is introduced first. Next, the details of joint prediction scheme are shown in the rest subsections.

A. Encoding Flow

The block diagram of the proposed stereo video encoder is shown in Fig. 2. The main differences between the left channel and the right channel are DE and mode pre-decision, which are introduced later. Note that reference frames from left and right channels are both reconstructed. After encoding, the left compressed data, M and L, and the right compressed data of a small amount, N and R, are transmitted.

B. Joint Block Compensation Scheme

In ME and DE steps of the right channel, the current block has two reference frames, as shown in Fig. 3. Grey region is

0-7803-8834-8/05/$20.00 ©2005 IEEE.

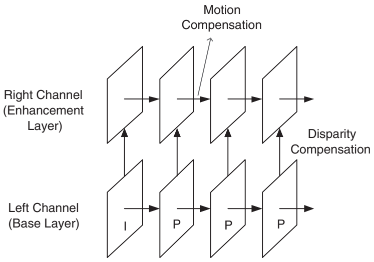

Fig. 1. Base-layer-enhancement-layer scheme of the proposed system.

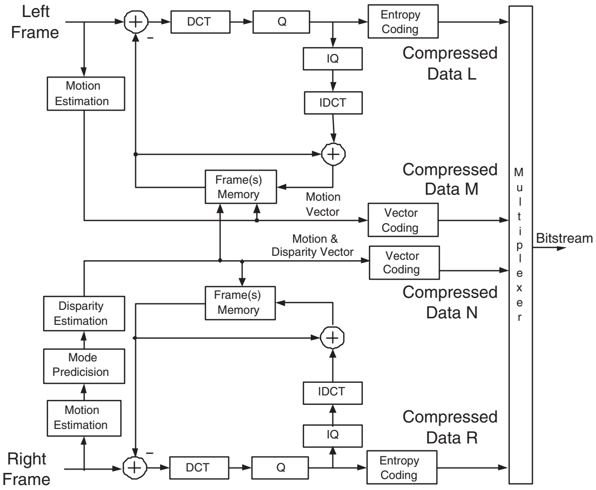

Fig. 2. Block diagram of the proposed stereo video encoder.

the search range of a reference frame. Note that search range of the left reference frame is not square because cameras are parallel-structured, so the candidate blocks are only on a belt of region.

There are three types of compensated blocks in the proposed stereo video encoder. 1) Motion-compensated block: it often occurs in the background due to its zero or slow motion. Occlusions between left and right frames will also compensated by this type of blocks. 2) Disparity-compensated block: 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 both foreground and background in it because the foreground and the background may be suitably predicted by different types of blocks.

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

$$D_{motion} = \min \sum_{\mathbf{B_R} \in SR_R(B)} |I_r(\mathbf{B}) - I_{r-1}(\mathbf{B_R})|, \quad (1)$$

$$D_{disparity} = \min \sum_{\mathbf{B_L} \in SR_L(B)} |I_r(\mathbf{B}) - I_l(\mathbf{B_L})|, \quad (2)$$

$$\begin{aligned} D_{j_n} = \textstyle\sum | I_r(B) - [W_n \cdot I_l(B'_L) + W_{n'} \cdot I_{r-1}(B'_R)] | \\ W_n + W_{n'} = I \end{aligned} , \quad (3)$$

where D_{motion} and $D_{disparity}$ are the minimum SAD values of motion- and disparity-compensated blocks, respectively. $I_r(\mathbf{B})$ is the right current block, $I_{r-1}(\mathbf{B_R})$ is the right reference block, and $I_l(\mathbf{B_L})$ is the left reference block. $SR_R(B)$ and $SR_L(B)$ are the search ranges in right and left reference frames of block B, respectively. DE and ME result in the best matching blocks, $I_l(B'_L)$ and $I_{r-1}(B'_R)$, in left and right reference frames. Then the proposed joint block is composed of the weighted sum of the two blocks, $I_l(B'_L)$ and $I_{r-1}(B'_R)$. W_n and $W_{n'}$ are complementary weighting functions that describe a weighting parameter or some example patterns such as A to D, as shown in Fig. 3. In (3), the SAD value D_{j_n} is derived. I indicates that weighting parameters all equals one after weighted sum. Finally, the mode decision is described as follows,

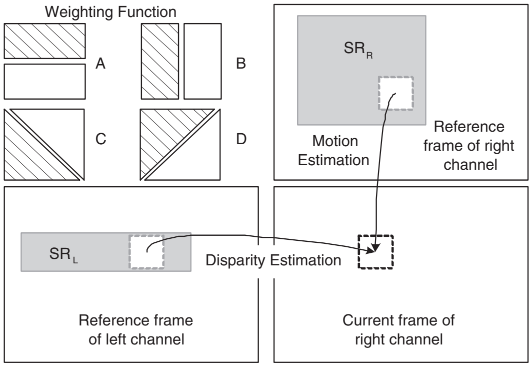

Fig. 3. The illustration of prediction directions and search range of two reference frames.

$$Mode = \arg \min_{mode} \{D_{motion}, D_{disparity}, D_{j_1}, ..., D_{j_n}\}. \quad (4)$$

C. Fast Algorithm of Motion Estimation in the Right Channel

In general stereo video systems, ME and DE are the key operations. However, compared with mono-video systems, additional ME and DE of the right channel greatly increase computational burdens. Therefore, in this and next sections, accurate motion vector prediction scheme and mode predecision scheme before DE are proposed. First, the correlation between motion vectors (MVs) and disparity vectors (DVs) is shown.

1) The correlation between DVs and MVs: The correlation between these four vectors Fig. 4 can be described as the following equation:

$$DV_{k-1} + MV_R = MV_L + DV_k \quad (5)$$

In general, because the difference between two frames in the temporal domain is tiny, we have the relation below,

$$DV_{k-1} \approx DV_k, MV_R \approx MV_L \quad (6)$$

According to the correlation, MVs derived in the left channel (MV_L) are set to be predictors of MVs in the right channel (MV_R). Because of the parallel camera structure, there is an global horizontal displacement between left and right channels. This displacement is called “global disparity.” In order to find the predictors, the global disparity should be derived first due to the relation between MVs and DVs introduced above. Here, we use a simple way to find the global disparity rather than global motion estimation (GME) scheme [6].

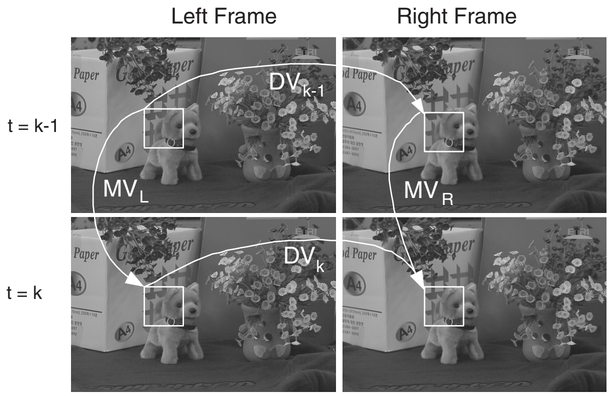

Fig. 4. The relation between DVs and MVs.

2) Background detection and global DE of the right channel: The background detection is determined as below,

$$F_{diff}(N) = \sum_{B} |I_r(\mathbf{B}) - I_{r-1}(\mathbf{B})|, \tag{7}$$

$$Background(N) = \begin{cases} true, & \text{if } MV_N(x,y) = (0,0) \text{ OR} \\ & \quad F_{diff}(N) < Threshold \\ false, & \text{otherwise.} \end{cases} \tag{8}$$

where $Background(N)$ is the state of the Nth block in the left frame. $MV_N(x,y)$ is the MV of the Nth block. If the MV is zero or $F_{diff}(N)$ is smaller than a threshold, this block is probably belongs to background, so $Background(N)$ is set to $true$. After this step, the global disparity vector, GD, is derived as follows,

$$GD = \arg\max_{DV}\{\ Num(\ DV\ )\ \} \tag{9}$$

The DVs of these zero motion blocks are gathered for statistics. The DV with the highest appearing frequency, $Num(DV)$, is regarded as GD. For the first P-frame in the left channel, background detection scheme is used to find GD. Before ME of the right channel, the corresponding block (in the left frame) of the current block (in the right frame) can be found by using GD. Then the MV of this related block is regarded as the predictor of the current block in the right frame, and MV_R can be derived correctly within only small search range to reduce computation. However, the DVs of background are usually smaller than those of foreground. If SAD is over a threshold, we adaptively extend the search range to find a better MV. Then, a more precise global disparity vector is fed back to the system. To avoid error propagation, GD can be updated every M frames, where M is a flexible parameter.

D. Mode Pre-decision Before Disparity Estimation

In our experiments shown in Fig. 5, statistics show that in a frame of the right channel, 40%–70% blocks are motion-compensated, 25%–60% blocks are joint-compensated, only about 5% blocks are disparity-compensated. Because a joint block results from both ME and DE, it means 25%–60% blocks need to perform DE and ME. From above analysis, over 95% blocks must perform ME, while only 40%–60% blocks must perform DE. Thus, unnecessary DE step should be skipped to reduce computational complexity. From our analysis, the MV-predicted blocks are often have zero motion, such as blocks in the background or blocks with slow motion caused by moving cameras. Therefore, by utilizing these features, mode pre-decision scheme is proposed by following equations,

$$SAD_{ME} = \min \sum_{\mathbf{B}' \in SR(B)} |I_r(\mathbf{B}) - I_{r-1}(\mathbf{B}')|, \tag{10}$$

$$Skip = \begin{cases} true, & \text{if } F_{diff} < Threshold_1 \quad \text{AND} \\ & \quad SAD_{ME} < Threshold_2 \\ false, & \text{otherwise.} \end{cases} \tag{11}$$

If $F_{diff} < Threshold_1$ and $SAD_{ME} < Threshold_2$ are simultaneously established, the mode of this block is usually MV-predicted. Then DE of this block is skipped, and the computational burdens is decreased.

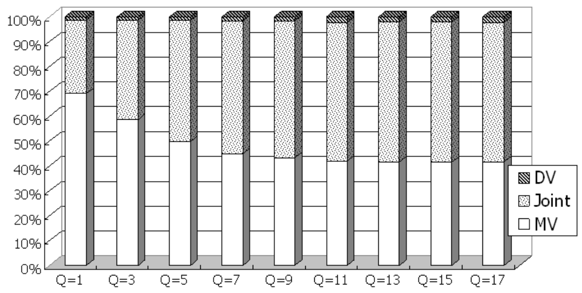

Fig. 5. Percentage of block prediction types of sequence“Race2.”

III. Experimental Results and Analysis

The proposed system is compared with MPEG-4 Simple Profile (SP) and Temporal Scalibility (TS) encoder [4]. Rate-distortion performance of only right channels (enhancement layer) are compared because the left channels are all encoded by MPEG-4 SP. Sequence “Race2” (320×240, 30 fps) and “Soccer2” (720×480, 30 fps) are taken as test sequences.

Figure 6 shows the comparison between the proposed algorithm, MPEG-4 TS, and MPEG-4 SP. Proposed joint prediction scheme is 2–3dB better than other two MPEG-4 profiles. Figure 7 shows the performance of different coding tools. It is shown that without joint block compensation, the PSNR degradation is serious. The proposed algorithms (curve “Joint Block” and “Joint Prediction”) is over 3dB better than MPEG-4 SP at low bit-rate, which is the target bit-rate for the right channel according to the “asymmetrical spatial resolution property” [8]. Besides, after applying MV-predictor prediction and mode pre-decision scheme (curve “Joint Prediction”), most of the computational complexity can be reduced, as shown in Fig. 8. In our experiments, the search range of ME is set to ±16 horizontal and vertical, and the search range of DE is set to ±32 horizontal and ±4 vertical. Without our algorithm, the overhead (additional search points relative to mono-video systems) is 150%. The proposed algorithm reduce the overhead under 20% in some cases, that is, it can speed up 8–9 times. From Fig. 9, we can see that if the search range is

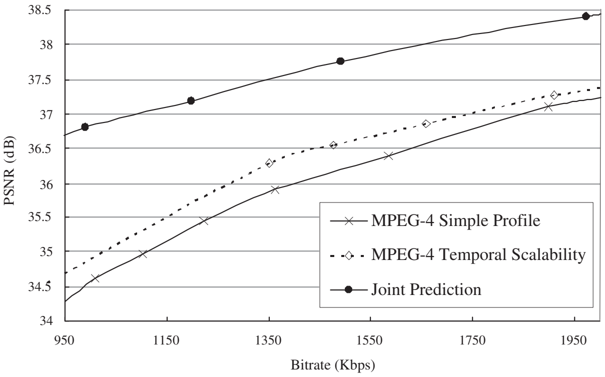

Fig. 6. Rate-distortion curve of sequence "Soccer2."

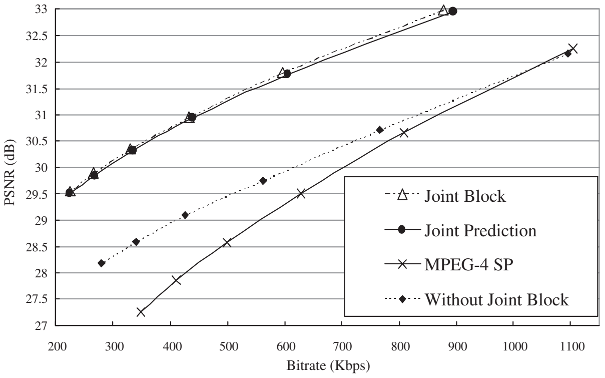

Fig. 7. Rate-distortion curve of sequence "Race2."

reduced from ±16 to ±2, the PSNR degradation is only about 0.1dB, while the computational complexity is greatly reduced. Figure 10 demonstrates the subjective quality of the proposed algorithm. The reconstructed frames of the proposed coding system with the bitrate 334.47 Kbps are shown in Fig. 10(a)(d). Figure 10(b)(e) is the proposed system without MV predictor prediction and mode pre-decision scheme. It shows that it is hard to recognize the difference between them. However, block artifacts can be easily observed in Fig. 10(c) and (f), which are encoded by MPEG-4 SP.

IV. Conclusion

In this paper, we propose a stereo video coding system with joint prediction scheme which combines three coding tools. Joint block compensation utilizes the weighted sum of motion- and disparity-compensated blocks. Our system outperforms MPEG-4 TS and SP by 2–3 dB in rate-distortion performance. Moreover, MV-prediction and mode pre-decision utilize the correlation between MVs and DVs, which reduce the computational burdens with up to 8–9 times acceleration. This system can be easily integrated by the existing standard and extend to the multi-view video coding system.

Algorithm	Race2	Flamenco	Soccer2	Puppy	Golf
Full Search (L)	1024	1024	1024	1024	1024
Full Search (R)	1536	1536	1536	1536	1536
Joint Prediction	381.42	200.45	415.75	177.62	234.89
Search point reduction	75.20%	86.95%	73.92%	88.44%	84.71%
Overhead (original)	150.00%	150.00%	150.00%	150.00%	150.00%
Overhead (proposed)	37.24%	19.58%	40.60%	17.35%	22.94%
Speed Up	4.03	7.66	3.69	8.50	8.65

Fig. 8. Search points and reduction rate of proposed scheme with and without complexity-reduction scheme.

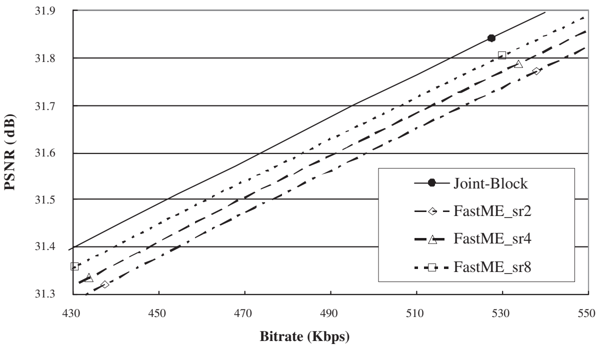

Fig. 9. Rate-distortion curve of Fast algorithm with various search range of sequence "Race2."

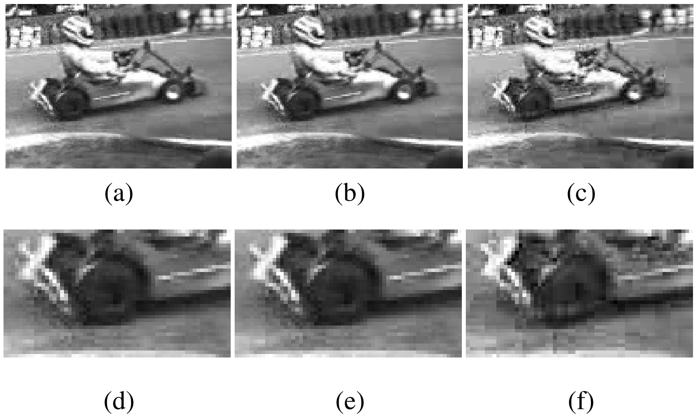

Fig. 10. Reconstructed frames of frame 46 of the sequence "Race2." (a)Proposed Joint prediction scheme@334.47 Kbps. (b) Joint block compensation@330.16 Kbps. (c) MPEG-4 SP@348.84 Kbps. (d), (e), and (f) are zoom-in views of (a), (b), and (c), respectively

References

[1] 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.

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

[3] 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.

[4] *Proposed draft amendament No. 3 to 13818-2 (multi-view profile)*. ISO/IEC JTC 1/SC 29/WG11 N1088, 1995.

[5] J.-R. Ohm and K. Müller, "Incomplete 3-D multiview representation of video objects," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 9, no. 2, pp. 389–400, Mar. 1999.

[6] R.-S. Wang and Y. Wang, "Multiview video sequence analysis, compression, and virtual viewpoint synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 397–410, Apr. 2000.

[7] H.-Z. Jia, W. Gao, and Y. Lu, "Stereoscopic video coding based on global displacement compensated prediction," in *International Conference on Information and Communications Security*, 2003, pp. 61–65.

[8] S. Pastoor, "3D-television: a survey of recent research results on subjective requirements," *Signal Processing: Image Communication*, vol. 4, no. 1, pp. 21–32, 1991.

[9] S. Cho, K. Yun, B. Bae, Y. Hahm, C. Ahn, Y. Kim, K. Sohn, and Y. h. Kim, *Report for EE3 in MPEG 3DAV*. ISO/IEC JTC1/SC29/WG11 M9186, December 2002.