FAST MOTION ESTIMATION WITH INTER-VIEW MOTION VECTOR PREDICTION FOR STEREO AND MULTIVIEW VIDEO CODING

Li-Fu Ding, Pei-Kuei Tsung, Wei-Yin Chen, Shao-Yi Chien, and Liang-Gee Chen

DSP/IC Design Lab, Graduate Institute of Electronics Engineering, Department of Electrical
Engineering, National Taiwan University 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan,
Email: {lifu, iceworm, cvictor, shaoyi, lgchen}@video.ee.ntu.edu.tw

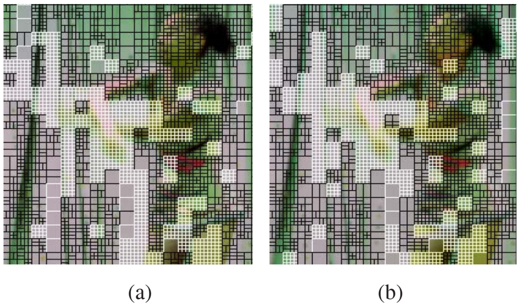
ABSTRACT

3-D video will become one of the most important video technologies in the next generation of television. Due to ultra high data bandwidth requirement for 3-D video, effective compression technology becomes an essential part in the infrastructure. Thus stereo and multiview video coding (MVC) plays a critical role. However, MVC systems require much more computational complexity relative to mono-view video coding systems. Therefore, an efficient prediction scheme is necessary for encoding. In this paper, a new fast motion estimation (ME) algorithm is proposed. By utilizing disparity estimation (DE) to find corresponding blocks between different views, the coding information such as motion vectors can be effectively shared and reused from the coded view channel. Therefore, the computation for ME in most view channels can be greatly reduced. Experimental results show that compared with the full search block matching algorithm applied to both ME and DE, the proposed algorithm saves 95% computation with near-FSBMA quality.

***Index Terms*—** 3D-video, MVC, video coding, H.264/AVC

1. INTRODUCTION

Multiview video can provide users with a sense of complete scene perception by transmitting several views to the receivers simultaneously. It can give users a vivid information about the scene structure. Moreover, it can also provide the capability of 3D perception by respectively showing two of these frames to the eyes. With the technology of 3D-TV [1] and free viewpoint TV (FTV) [2] getting more and more mature, multiview video coding (MVC) draws more and more attention. In recent years, JVT/MPEG 3D auido/video (3DAV) group has worked toward the standardization for MVC [3], which also advances the multiview video applications. From the discussion in JVT/MPEG 3DAV meetings, the developed coding scheme for multiview video settings mainly uses H.264/AVC with exploiting temporal and inter-view dependencies [4]. That is, many coding tools of MVC in the related research area are based on the hybrid coding scheme and highly related to H.264/AVC.



(a) (b)

Fig. 1. Illustration of MB partition after variable-block-size ME. Two views are independently encoded without DE. (a) Left view. (b) Right view.

Although MVC is an emerging technology, huge amount of video data and ultra high computational complexity make it difficult to be realized. An H.264/AVC encoder requires computing power of about 1.3 tera-operations/second (TOPS) on a general-purpose processor to encode single-view HDTV720p videos (1280×720, 30 frames/second) in real time [5]. Different from mono-view video coding, disparity estimation (DE) is also utilized to reduce inter-view redundancy in MVC. Taking coding efficiency into consideration, block-based DE, like motion estimation (ME), is more suitable for MVC because it has better compatibility with the existing video coding standards. Consequently, the prediction part, which consists of ME and DE, becomes the most computationally intensive part in a MVC system.

In a MVC system, ME removes the temporal redundancy while DE removes the inter-view redundancy. Because of the setup structure of multiple cameras, there is close relation between motion vectors (MVs) and disparity vectors (DVs) in neighboring frames. [6] By utilizing the correlation between motion and disparity fields, some new coding methods for MVC have been proposed [6][7] to save coding bits for residue or MVs, then the coding performance can be enhanced. On the other hand, according to the inter-view correlation, another kind of redundancy called "computational redundancy" exists in addition to temporal and inter-view re-

1-4244-1484-9/08/$25.00 ©2008 IEEE

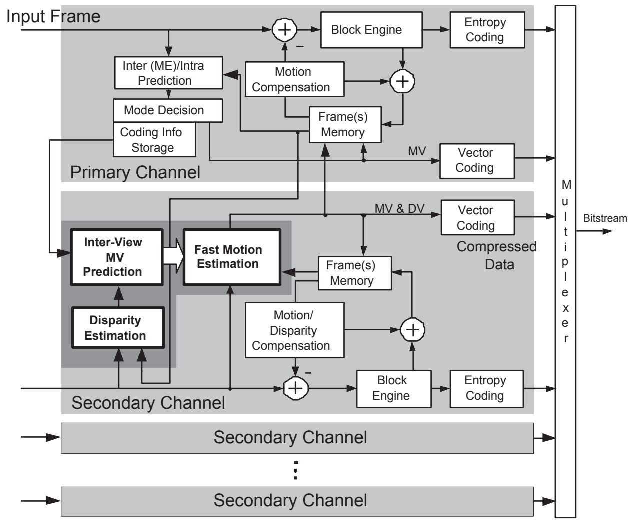


Fig. 2. Block diagram of the proposed multiview video encoder.

dundancies. Based on this concept, a fast prediction algorithm has been proposed to save the computation of ME for stereo video coding in our previous work [8]. However, the coding structures in MVC are more complex than that in stereo video coding. Besides, the previous work cannot deal with variable-block-size ME and complex mode decision.

It is also observed that the mode distribution of two views are very similar. On condition that cameras are setup with close parallelized structure, the video contents of different views are usually similar. The view similarities exist not only in the video contents but also in the prediction modes. An example is shown in Fig. 1. Two views are encoded separately by an H.264/AVC encoder. The macroblock (MB) partition is marked on the reconstructed frames. Black and white blocks represent inter- and intra-predicted blocks respectively. It shows that the inter-view correlation is high. In other words, if the correlation is effectively exploited, the computational complexity can be greatly reduced.

In this paper, an new fast ME algorithm is proposed for MVC. Based on the fact that the video contents are highly related between view channels, the proposed algorithm greatly reduces computational complexity while maintains video quality. The remainder of the paper is organized as follows. The proposed MVC system architecture is first introduced in Section 2. Then the proposed algorithm is presented in Section 3. Section 4 shows the simulation results. Finally, Section 5 concludes this paper.

2. SYSTEM ARCHITECTURE

The block diagram of the proposed multiview video encoder is shown in Fig. 2. The encoder adopts the coding tools defined in H.264/AVC standard. Input views are classified into two types of view channels, the primary channel and the secondary channel. A view channel is regarded as a primary channel if no reconstructed frames in other view channels are used for reference when performing mode decision. Therefore, there are no DE operations in primary channels. The coding flow of a primary channel is identical to the flow of mono-view video coding. The block engine includes quantization, transform, and deblocking filter, etc.. After the Lagrange mode decision, the coding information, including MB partition and the corresponding MVs, are stored in the encoder.

The main difference between primary and secondary channels is the dark grey part, which contains DE, inter-view MV prediction, and fast ME. Each of them is introduced in the following subsections. In the proposed encoder, DE is performed prior to ME. The purpose of performing DE first is to extract the correlation between views. Therefore, the coding information of the neighboring coded view can be derived after DE. Then the inter-view MV prediction part decides an initial guess for each MB partition for the current MB. Proposed fast ME is a predictor-based ME algorithm. The number of primary and secondary channels depends on the coding structure. The number of secondary channels is usually much more because DE can effectively improve coding efficiency [4]. After all views are encoded, the compressed bitstream of each channel is assembled and transmitted.

3. PROPOSED FAST MOTION ESTIMATION WITH INTER-VIEW MOTION VECTOR PRECITON

There are seven MB partition types according to their block size such as 16×16, 8×8, and 4×4, etc. in H.264/AVC. After DE and intra prediction in a secondary, inter-view MV prediction is proposed to provide an initial guess of MV for each MB partition for the current MB. The illustration of inter-view MV prediction is shown in Fig. 3. After the frame in the primary channel is encoded, the coding information is stored in the memory. The location of the disparity-compensated block has already been derived from DE shown as the grey area in Fig. 3 (a). The grey area is split into sixteen 4×4 sub-blocks. Each sub-block covers a 4×4 area in the reference frame. Then each sub-block is assigned a MV which is the same as the MV in the 4×4 area which is covered by the disparity-compensated block in the coded reference frame. Note that if the 4×4 area contains more than one MV, the assigned MV is the MV of the coded sub-MB with the largest overlapped area by the disparity compensated block. To prevent the prediction error propagation, there is not any early termination and fast prediction scheme applied in the primary channel. It means all kinds of cost must be calculated in the primary channel. No matter what kind of mode is selected for coding, the best inter prediction mode and its corresponding MVs are stored. Therefore, if the covered MB in the reference frame is predicted by intra or skip mode, the MB partition and its

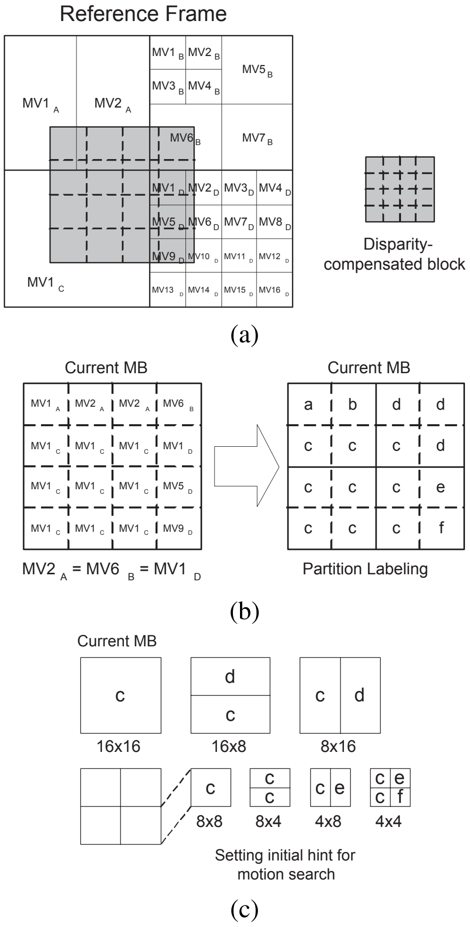


Fig. 3. Illustration of inter-view MV prediction.

corresponding MVs can be still adopted for the proposed algorithm. After each 4×4 sub-block is assigned a MV, the process of “partition labelling” begins. The sub-blocks with the same MVs are assigned to the same label, as shown in Fig. 3 (b). Then, an initial guess of the MV is set for each MB type according to partition labelling. An example is shown in Fig. 3 (c). To provide a good initial guess of a MV, the most representative value should be chosen. When setting the initial guess of MV for a 16×16 block, the value of the label which appears the most times is chosen as an initial guess. If there is not any label which appears the most times, the top-left label is set as the initial guess. In the case of Fig. 3 (c), although the 8×8 and two 8×4 sub-blocks choose the same label as the initial guess, the rate-distortion cost of two 8×4 sub-blocks is obviously larger than that of one 8×8 sub-block due to the penalty for MV coding bits.

In addition to the initial guess from inter-view MV prediction, the MVs of the left, top, and top-right neighboring MBs, and zero MV, are also adopted as initial guesses to enhance the coding efficiency. That is, for each MB type, there are five initial guesses of MVs. The optimum initial guess is chosen by Lagrange mode decision. Then the refinement with small search range is performed around the optimum initial guess.

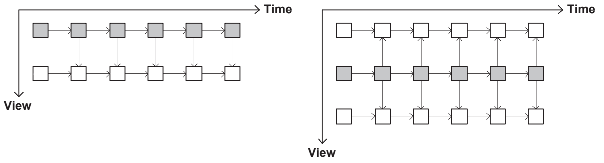


Fig. 4. Two-view and three-view coding structures for experiments. The grey blocks represent the frames in the primary channel, and the white blocks represent the frames in the secondary channels.

Sequence	Akko&Kayo	Rena	Ballroom	Exit
ME Complexity Reduction	95%	95%	95%	95%
Quality drop (dB)	0.06 dB	0.01 dB	0.05 dB	0.1 dB
Inter-view MV predictor	81.7%	94.9%	83.3%	89.0%
Left predictor	10.6%	4.6%	11.1%	8.8%
Top predictor	3.5%	0.3%	2.1%	1.2%
Top-right predictor 3	1.3%	0.1%	0.9%	0.4%
Zero predictor	2.8%	0.1%	2.7%	0.6%

Table 1. ME complexity reduction in a secondary channel and percentage of five initial guesses of MVs.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed algorithm is implemented by modifying the MVC-configuration in JSVM4.5 [9]. It is compared with the multicast coding, in which the full search block matching algorithm (FSBMA) is applied to both ME and DE in the coding structures. Rate-distortion performance of only secondary channels are compared because the ME parts in the primary channels in both cases are implemented with FSBMA. Sequences “Akko&Kayo,” “Ballroom,” “Exit,” “Ballroom,” and “Rena,” with size 640×480 are tested. They are standard sequences released by JVT/MPEG 3DAV Group [4]. Two and three view channels of these sequences are chosen for simulation. The illustrations of the coding structures are shown in Fig. 4. The grey blocks represent the frames in the primary channel, and the white blocks represent the frames in the secondary channels. The search ranges of DE and ME are both [-32, +31] in horizontal and vertical directions.

Figure 5 and Fig. 6 show the rate-distortion curves of two test sequences. The proposed algorithm with different refinement ranges is compared with FSBMA. It is shown that there is almost no quality difference between them. The distribution of MV difference, that is, the difference between the proposed initial guess and the final MV, is shown in Fig. 7. It is observed that almost all MV differences are located within two pixels. It shows that the proposed inter-view MV predic-

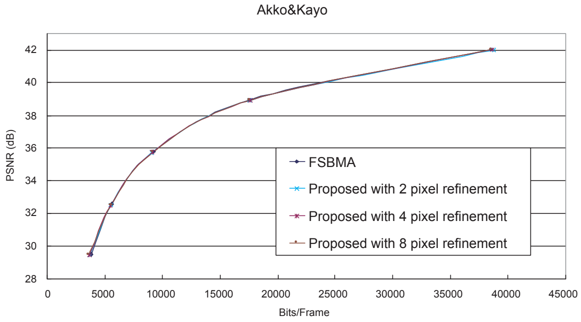


Fig. 5. Rate-distortion comparison between FSBMA and proposed fast ME with different refine ranges. “Akko&Kayo.”

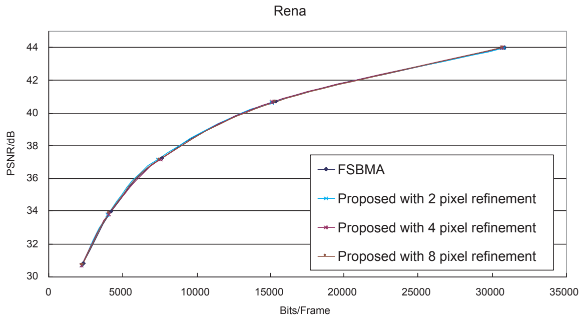


Fig. 6. Rate-distortion comparison between FSBMA and proposed fast ME with different refine ranges. “Rena.”

tion is accurate. Therefore, the refinement range [-2,+2] in both horizontal and vertical directions are sufficient.

Talble 1 shows ME complexity reduction and the percentages five initial guesses which are adopted for further refinement to get final MVs. Compared with FSBMA, ME complexity of a secondary channel is greatly reduce because the search candidates is only about 5% of search candidates in FSBMA. Meanwhile, the quality can still be maintained. Note that if the search range of the primary channel gets larger, the search candidates of secondary channels remain the same because the proposed algorithm is independent of the size of search range. The quality degradation is within 0.1 dB. On the other hand, it also is observed that the initial guess of MV proved by inter-view MV prediction is the most accurate and provide lower rate-distortion cost than other predictor types. Therefore, the proposed fast ME with inter-view MV prediction utilizes the inter-view correlation successfully.

5. CONCLUSION

In this paper, we propose a fast ME algorithm with inter-view MV prediction for stereo and multiview video coding. An efficient system architecture for MVC which contains primary and secondary view channels are also presented. By utilizing the coding information of a coded primary channel, 95% ME complexity of secondary channels can be effectively reduced with only 0.01–0.1 dB quality degradation. Therefore, the proposed algorithm effectively exploit the correlation between views.

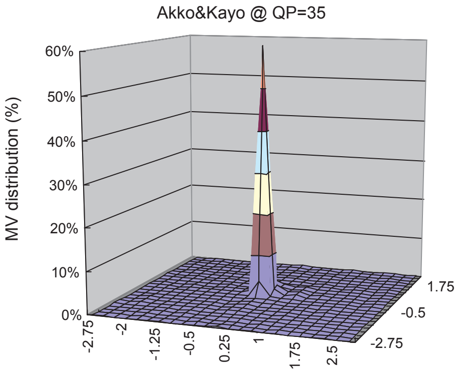


Fig. 7. Distribution of differences between initial guesses and final MVs.

6. REFERENCES

[1] F. Isgrò, E. Trucco, P. Kauff, and O. Schreer, “Three-dimensional image processing in the future of immersive media,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 388–303, Mar. 2003.

[2] M. Tanimoto, “Free viewpoint television - FTV,” in *Proceedings of 2004 Picture Coding Symposium*, Dec. 2004.

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

[4] ISO/IEC JTC1/SC29/WG11 N6501 W8019, *Description of Core Experiments in MVC*, Apr. 2006.

[5] Y.-W. Huang, T.-C. Chen, C.-H. Tsai, C.-Y. Chen, T.-W. Chen, C.-S. Chen, C.-F. Shen, S.-Y. Ma, T.-C. Wang, B.-Y. Hsieh, H.-C. Fang, and L.-G. Chen, “A 1.3TOPS H.264/AVC single-chip encoder for HDTV applications,” in *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2005.

[6] Y. Luo, Z. Zhang, and P. An, “Stereo video coding based on frame estimation and interpolation,” *IEEE Trans. Broadcast.*, vol. 49, no. 1, pp. 14–21, Jan. 2003.

[7] X. Guo, Y. Lu, and W. Gao, “Inter-view direct mode for multi-view video coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 12, pp. 1527–1532, Dec. 2006.

[8] L.-F. Ding, S.-Y. Chien, and L.-G. Chen, “Joint prediction algorithm and architecture for stereo video hybrid coding systems,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1324–1337, Nov. 2006.

[9] ISO/IEC JTC1/SC29/WG11 N6501 N7829, *AHG on Multiview Video Coding*, ISO/IEC, Jan. 2006.