IEICE TRANS. INF. & SYST., VOL.E94–D, NO.7 JULY 2011

LETTER

Compatible Stereo Video Coding with Adaptive Prediction Structure

Lili MENG[†a)], **Yao ZHAO**[†], *Nonmembers*, **Anhong WANG**[††], **Jeng-Shyang PAN**[†††], *Members*, *and* **Huihui BAI**[†], *Nonmember*

SUMMARY A stereo video coding scheme which is compatible with monoview-processor is presented in this paper. At the same time, this paper proposes an adaptive prediction structure which can make different prediction modes to be applied to different groups of picture (GOPs) according to temporal correlations and interview correlations to improve the coding efficiency. Moreover, the most advanced video coding standard H.264 is used conveniently for maximize the coding efficiency in this paper. Finally, the effectiveness of the proposed scheme is verified by extensive experimental results.

key words: *stereo video coding, compatible, temporal correlation, interview correlation, adaptive prediction structure*

1. Introduction

3D video can provide more vivid and accurate information than monoview video and the technologies of 3DTV become more and more mature [1]. With the increasing demands for realistic multimedia contents, 3D video technologies have drawn interest from both academia and industry lately [2], [3]. So far, 3D video has been used widely in many applications, such as, 3D telemedicine, 3D visual communications, 3DTV and virtual reality [4], [5]. Among the various 3D video representations, stereo video is the most widely used because of its simple format [6]. Stereo video only includes left video and right video which are captured by two cameras. Although stereo video is attractive, the amount of the stereo video data and the computational complexity are at least twice those of monoview video [7]. It's difficult to store and transmit stereo video data. Therefore, the compression of stereo video data is necessary.

In order to compress stereo video efficiently, the interview correlations which exist in the different views at the same time instant and the temporal correlations which exist in the same view at different time instants should be exploited efficiently. Motion compensation prediction (MCP) and disparity compensation prediction (DCP) can utilize temporal correlations and interview correlations respectively [5]. At present, there are two kinds of prediction technologies. One is simulcast coding where only MCP is used. In the simulcast coding, there are two encoders. The other is joint prediction technologies [5], [7], which have been developed well. In the joint prediction technologies, MCP and DCP are used to reduce the temporal redundancy and spatial redundancy of the right video. But for the left video, only temporal correlations are exploited. The processor which compresses monoview video has been applied widely up to now. Using the above two coding technologies, the monoview-processor can't compress stereo video conveniently. In this paper, we propose a scheme which is not only compatible with monoview-processor but also can exploit temporal correlations and interview correlations efficiently. In addition, for getting good coding efficiency, the most advanced video coding standard H.264 is imposed.

This paper proposes a scheme which is based on H.264 and compatible with present monoview-encoder. An adaptive prediction structure is also proposed in this paper. Considering real-time nature of the proposed scheme, group of picture (GOP) is defined. In this paper, the length of GOP is four and there are four kinds of compatible flexible prediction modes which have been proposed in our previous work [8]. The proposed stereo video coding scheme can choose the prediction mode according to the temporal correlations and interview correlations. For decreasing the computational complexity, only the correlations of low-frequency sub-band are utilized. In the adaptive prediction structure, temporal correlations and interview correlations are exploited efficiently to get better coding efficiency.

The rest of this paper is organized as follows. Section 2 describes the compatible stereo video coding with adaptive prediction structure. The experimental results are given in Sect. 3. Finally, Sect. 4 concludes this paper.

2. Compatible Stereo Video Coding with Adaptive Prediction Structure

2.1 Adaptive Prediction Structure

In our previous work [8], the flexible prediction modes of stereo video coding have been proposed. Take the length of GOP is 4 as an example and Fig. 1 describes it. In the Fig. 1, L_{2i-1} denotes the left frame at the time of 2i-1. L_{2i} denotes the left frame at the time of 2i. R_{2i-1} denotes the right frame at the time of 2i-1. In the same way, R_{2i} denotes the right frame at the time of 2i. From Fig. 1, we can see

Manuscript received July 6, 2010.
Manuscript revised March 22, 2011.

[†]The authors are with Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China.
[††]The author is with Taiyuan University of Science and Technology, Taiyuan, 030024, China.
[†††]The author is with the Department of Electronic Engineering, Kaohsiung University of Applied Sciences, Kaohsiung, 807, Taiwan.
a) E-mail: 05120377@bjtu.edu.cn

DOI: 10.1587/transinf.E94.D.1506

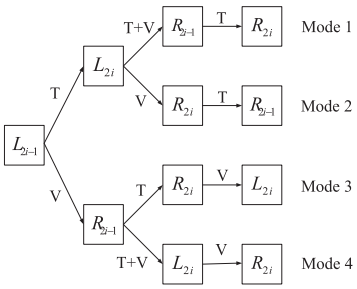


Fig. 1 Flexible prediction modes of stereo video coding with GOP=4.

that there are four kinds of prediction structures according to the exploitation of temporal correlations and interview correlations.

However, which mode is suitable for different GOPs of different stereo video sequences? In this subsection, we propose an adaptive mode conversion method which can solve this problem and is described by Fig. 2.

In order to reduce the computational complexity, we only compute the correlations of low-frequency sub-band whose temporal correlations and interview correlations can express the correlations of the original video. In this paper, the low-frequency sub-band is expressed by sub-band-LL. Figure 3 depicts the process of adaptive mode conversion. Firstly, the stereo video is processed by DWT. Then we extract the low-frequency sub-band. The sub-band-LL is processed by the mode-adjudicate which is shown in Fig. 3. The sign bits of mode (S_m) are achieved and transmitted to the mode-conversion. After mode-conversion, the conversed video of mode i which is determined by S_m is received.

In the Fig. 3, F_1 denotes the first frame of the current mode-adjudicate. $LL_{l,2i-1}$ indicates the low-frequency sub-band of the frame-L_{2i-1}. In the same way, $LL_{l,2i}$ and $LL_{r,2i-1}$ represent the low-frequency sub-band of frame-L_{2i} and frame-R_{2i-1} respectively. T_c and V_c denote the temporal correlation and interview correlation respectively. Take the choice of second frame as an example, Fig. 4 depicts the temporal correlation-T_c and interview correlation-V_c. T_c and V_c are computed as:

$$T_c = 10\lg \frac{255^2 \times row \times col}{\sum_{i=1}^{row}\sum_{j=1}^{col}\left(LL^c_{l,2i}(i,j) - LL_{l,2i}(i,j)\right)^2}$$

$$V_c = 10\lg \frac{255^2 \times row \times col}{\sum_{i=1}^{row}\sum_{j=1}^{col}\left(LL^c_{r,2i-1}(i,j) - LL_{r,2i-1}(i,j)\right)^2} \quad (1)$$

Where row and col are the height and width of the low-frequency sub-band. $LL^c_{l,2i}$ and $LL^c_{r,2i-1}$ denote the compensation of $LL_{l,2i}$ and $LL_{r,2i-1}$ by the first frame $LL_{l,2i-1}$ respectively.

The process of mode-adjudicate is as following:

(1) $LL_{l,2i-1}$ is as the first frame.

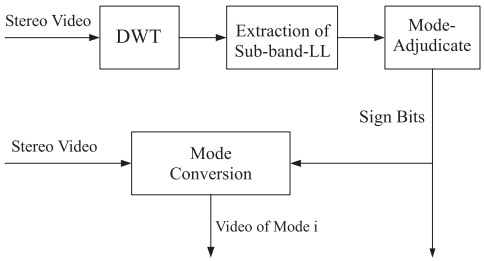


Fig. 2 Adaptive mode conversion.

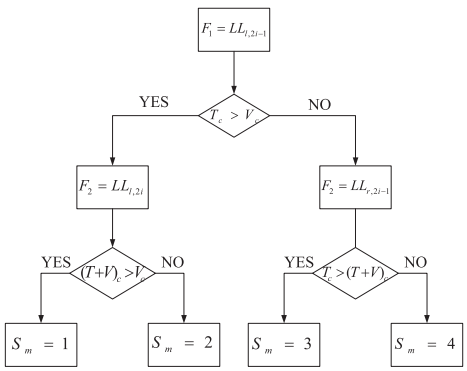


Fig. 3 The structure of mode-adjudicate.

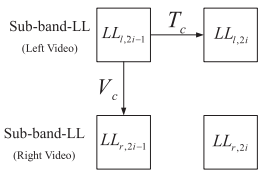


Fig. 4 The picture of temporal correlation and interview correlation.

(2) Judging which low-frequency sub-band is as the second frame:
If $T_c > V_c$ which express temporal correlation is stronger than interview correlation, $LL_{l,2i}$ is the second frame;
else $LL_{r,2i-1}$ is the second frame which denotes the interview correlation is intenser than temporal correlation.

(3) When $LL_{l,2i}$ is the second frame, if $(T+V)_c > V_c$, $S_m = 1$; else $S_m = 2$;
When $LL_{r,2i-1}$ is the second frame, if $T_c > (T+V)_c$, $S_m = 3$; else $S_m = 4$.

After mode-adjudicate, we obtain the sign bits of mode (S_m). According to sign bits (S_m), the mode-conversion is performed. If S_m is i ($i = 1, 2, 3, 4$), the video of mode i is got. It is obvious that the prediction mode of every GOP isn't fixed. Mode 1 and mode 2 are suit for the stereo video whose temporal correlations are stronger than interview correlations. If interview correlations are intenser than temporal correlations, mode 3 and mode 4 are more conformable. The proposed adaptive mode conversion which make the prediction structure is fittest can regroup every GOP. The

regroup GOP structure can exploit the temporal correlations and the interview correlations effectively.

2.2 Compatible Stereo Video Coding

Traditional stereo video coding based on H.264 is that left video sequence and right video sequence are encoded by H.264 respectively. Figure 5 shows the traditional stereo video coding scheme based on H.264. The processor which compresses the monoview video has been used widely in every field at present. From the Fig. 5, we can see that the traditional stereo video coding needs two encoders and only temporal correlations are used. Therefore, it is obvious that the scheme of traditional stereo video coding isn't compatible with the encoder of monoview video and interview correlations aren't exploited. In this paper, a compatible stereo video coding with adaptive prediction structure is proposed. The adaptive prediction structure which has been depicted in the above subsection can impose temporal correlations and interview correlations efficiently.

The above Fig. 6 depicts the proposed compatible stereo video coding with adaptive prediction structure. In this proposed scheme, both the temporal correlations and interview correlations are exploited. The existing monoview video equipment can process stereo video straightway and the most advanced video standard H.264 is imposed conveniently. From Fig. 6, we can get the process of the proposed compatible stereo video coding. Firstly, the left video and right video are preprocessed by adaptive mode conversion which has been introduced in the above subsection. The converted video of mode i which has been described by Fig. 7 is obtained. At the same time, the sign bits (S_m) are received. After adaptive mode conversion, the video of mode i is encoded by H.264. Meanwhile, the sign bits are encoded by entropy encoder. The encoded bit streams are transmitted to the decoder. At the decoder, the compressed video bit streams and the compressed sign bits are decoded by H.264 decoder and entropy decoder respectively. At last, the reconstructed stereo video is achieved by reverse conversion with the help of sign bits (S_m).

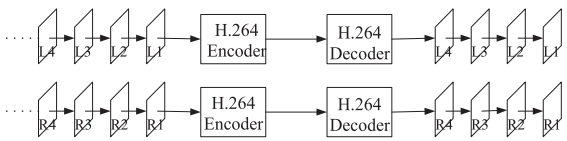


Fig. 5 Traditional stereo video coding based on H.264.

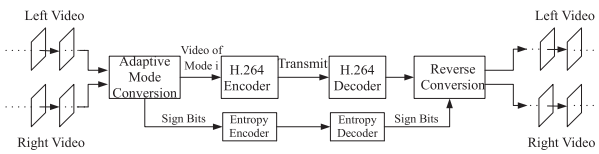


Fig. 6 Compatible stereo video coding based on H.264.

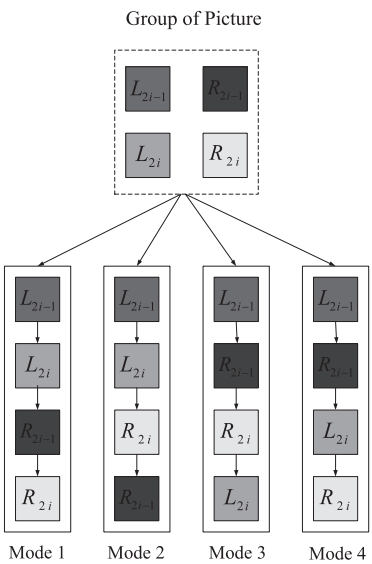


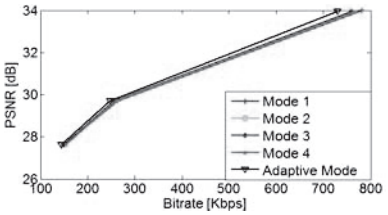
Fig. 7 The conversion video of all the mode.

3. Experimental Results

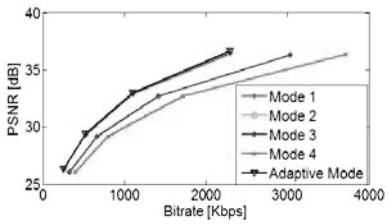
Experiments are performed on the stereo video sequences which are named "soccer2", "puppy", "soccer" and "rabbit". The resolution of tested stereo video is 720×480. The frame rate is 30. In this paper, we choose 120 frames of every video sequence to do the experiments. All the proposed prediction modes and the adaptive prediction structure are performed. In the adaptive prediction structure, the prediction mode of every GOP is determined by the temporal correlations and interview correlations. The computational complexity of temporal correlations and interview correlations is very large. For reducing the computational complexity, only the correlations of low-frequency sub-band are computed. The low-frequency sub-bands of the tested sequences are shown in Fig. 8. Both low-frequency sub-bands and high-frequency sub-bands can express the correlation of the original sequences. However, human is insensitive for the high-frequency and most energies are in low-frequency sub-bands. Therefore, the correlations of low-frequency sub-bands are exploited to reduce the computing complexity.

The comparison of the rate-distortion performances are shown in Fig. 9. From the Fig. 9, we can find that the cod-

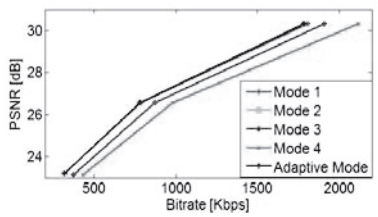
Fig. 8 The low-frequency sub-bands of tested sequences.



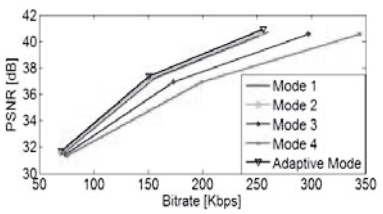
(a) Soccer2



(b) Puppy



(c) Soccer



(d) Rabbit

Fig. 9 The comparison of the rate-distortion performances.

ing efficiency of the proposed adaptive mode is better than the other prediction modes. In this paper, all the simulant schemes are compatible with the existing monoview-processor. In a word, the proposed compatible stereo video coding with adaptive prediction structure is effective and can make the present monoview-processor compress stereo video expediently.

4. Conclusion

In this paper, a compatible stereo video coding with adaptive prediction structure is proposed. At present, the monoview video has been used widely and the equipment that processes the monoview video has been dominant. The prevenient schemes of stereo video coding aren't compatible with the present equipment which processes the monoview video. In this paper, a compatible scheme can make the present equipment compresses the stereo video conveniently. Meanwhile, an adaptive prediction structure is proposed to compress stereo video efficiently. In order to decrease the computational complexity of comparing temporal correlations and interview correlations, only the correlations of low-frequency sub-bands are considered. According to the temporal correlations and interview correlations of low-frequency sub-bands, the adaptive mode conversion can make the prediction of every GOP appropriate. Therefore, temporal correlations and interview correlations can be exploited efficiently. The experimental results demonstrate that the coding efficiency of proposed scheme is effective.

Acknowledgments

Special thanks to Huawei Tian and Gang Cao of Beijing Jiaotong University. This work was supported in part by National Natural Science Foundation of China (No.60776794, No.60903066, No.60972085), Sino-Singapore JRP (No.2010DFA11010), Beijing Natural Science Foundation (No.4102049), New teacher Foundation of State Education Ministry (No.20090009120006).

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.288–303, March 2004.

[2] R.-S. Wang, Multiview/stereoscopic video analysis, compression, and virtual viewpoint synthesis, PhD Thesis, Polytechnic, New York, June 1999.

[3] Il-L. Jung, T. Chung, K. Song, and C.S. Kim, "Efficient stereo video coding based on frame skipping for real-time mobile applications," IEEE Trans. Consum. Electron., vol.54, no.3, pp.1259–1266, Aug. 2008.

[4] L. Guan, S.Y. Kung, J. Larsen ed., Multimedia Image and Video Processing, CRC Press, Boca Raton, 2000.

[5] W. Yang, K. Ngan, J. Lim, and K. Sohn, "Joint motion and disparity fields estimation for stereoscopic video sequences," Signal Processing: Image Communicaiton 20, pp.265–276, Elsevier, 2005.

[6] P. Merkle, H. Brust, K. Dix, A. Smolic, and T. Wiegand, "Stereo video compression for mobile 3D services," 3DTV Conference, 2009.

[7] 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.

[8] L. Meng, Y. Zhao, J. Pan, and H. Bai, "Compatible stereo video coding with flexible prediction mode," PCSPA Conference, pp.755–758, Harbin, China, Sept. 2010.