Side Information Generation for Multi-view Distributed Video Coding Using Hybrid Search and Multi-selection

Zhenhua Tang[1, a], Bibo Huang[1, b], Xuebing Pei[2, c] , Tuanfa Qin[1, d] and Xufang Huang[1, e]

[1]School of Computer and Electronic Information, Guangxi University,

Nanning, Guangxi 530004, P.R. China

[2]China Ship Development and Design Center

Wuhan, Hubei 430064 P.R. China

[a]tangedward@126.com,[b]huangbibomc@163.com,[c]peixbhust@163.com,[d]tfqin@gxu.edu.cn, [e]hxf_andalan@163.com

Keywords: multi-view; distributed video coding; side information generation; hybrid search; multi-selection**.**

Abstract. The quality of side information (SI) has a direct impact on the rate-distortion (R-D) performance of multi-view distributed video coding (MDVC). In this paper, we propose an inter-view SI generation scheme for MDVC, which combines hybrid search and multi-selection. In the presented scheme, both feature matching and epipolar line search are employed to improve the accuracy of frame matching. Moreover, multi-selection has been used to obtain the optimal view disparity vectors (DVs). Experimental results show that under the same coding bitrates, the reconstructed video quality of the proposed scheme is superior to that of the method employing disparity compensated view prediction (DCVP) technique significantly. Additionally, without knowing the camera intrinsic parameters, the presented scheme can achieve the same decoded video quality compared with the approach using disparity based view synthesis (DBVS) technique.

1. Introduction

In past ten years, a multitude of studies have been concentrated on distributed video coding (DVC) due to the emergence of applications that require low complexity encoder, such as wireless multimedia sensor networks (WMSN), mobile video camera, and wireless video surveillance. Unlike the traditional video coding methods, such as MPEG-X and H.26X, the paradigms of DVC move the computational complexity from the encoder to the decoder, in terms of the Slepian-Wolf [1] and Wyner-Ziv [2] (WZ) theorems. Among DVC system, input video frames are categorized into two types: key frames and WZ frames. Key frames are encoded and decoded using traditional coding approaches, while channel coding techniques have been employed for coding WZ frames. In order to obtain good compression performance, only parity bits of WZ frames are transmitted from the encoder to the decoder. Due to absence of original WZ frames at the decoder, corresponding prediction version called side information (SI) must be created [3]. Hence, accurate SI may have a significant impact on the performance of the total system, such as rate-distortion (R-D) performance.

A large number of techniques focus on SI generation for single view DVC have been explored [4]. Typical methods, such as motion compensated temporal interpolation (MCTI) and its extension works [5-7], apply interpolation to the temporal adjacent two frames of WZ frame and then perform motion prediction to obtain corresponding SI frames. On the other hand, disparity compensated view prediction (DCVP) [8] is by far the most widely used approach to generate inter-view SI for multi-view DVC (MDVC) [9]. However, this algorithm may fail to determine the position of the same object in different views due to varying illumination and parallax. Hui et.al [10,11] developed an inter-view SI creation method, in which the feature information of original WZ frames at the encoder are extracted and transmitted to the decoder to improve the accuracy of frame matching in left and right views. But this would lead to high computational complexity at the encoder. Artigas et.al [12] proposed a spatial SI generation approach that combines temporal and spatial searching, called multi-view motion estimation (MVME). Unfortunately, any mistake in motion matching would be likely to affect the final quality of SI. C. Brites et.al [13] presented an inter-view SI generation scheme that adopts disparity based view synthesis (DBVS) techniques. This scheme exploits inter-view correlation by using epipolar line search techniques, and then realizes the final spatial orientation utilizing the camera intrinsic parameters. However, since the camera intrinsic parameters must be determined before zooming, the scheme cannot be applied to the cases that cameras are capable of zoom.

In this paper, we propose an inter-view SI generation scheme for MDVC, which combines hybrid search and multi-selection. In the presented scheme, both feature matching and epipolar line searching are employed to improve the accuracy of image matching. Furthermore, multi-selection has been used to achieve the optimal view disparity. We emphasize that the proposed scheme does not need to determine the camera intrinsic parameters beforehand, thus it can be used for applications that cameras are capable of zooming.

The rest of paper is organized as follows. Section 2 presents the system framework of MDVC adopted in the paper. Section 3 describes the proposed inter-view SI generation scheme. The experiment results and discussions are presented in section 4. Finally, section 5 concludes this paper.

2. System Overview

We consider a general framework of the three-view WZ distributed video codec based on discrete cosine transform (DCT) domain similar to [13], illustrated in Fig.1. However, we apply different methods to generate the inter-view SI, which will be described in section 3 in detail. At the encoder, videos captured by different cameras are encoded individually; while joint decoding is perform at the decoder. Specially, decoded information from the left and right views will provide auxiliary information for decoding the central view. And conventional video coding methods, such as MEPG-x and H.26L are used to code the contents captured by the left and right cameras. On the other hand, WZ video coding techniques are applied to the central view, in which key frames are coded by JPEG and low density parity check code (LDPC) is used to code WZ frames. When decoding the current WZ frame in the central view, two types of SI frames including temporal and inter-view SI must be generated.

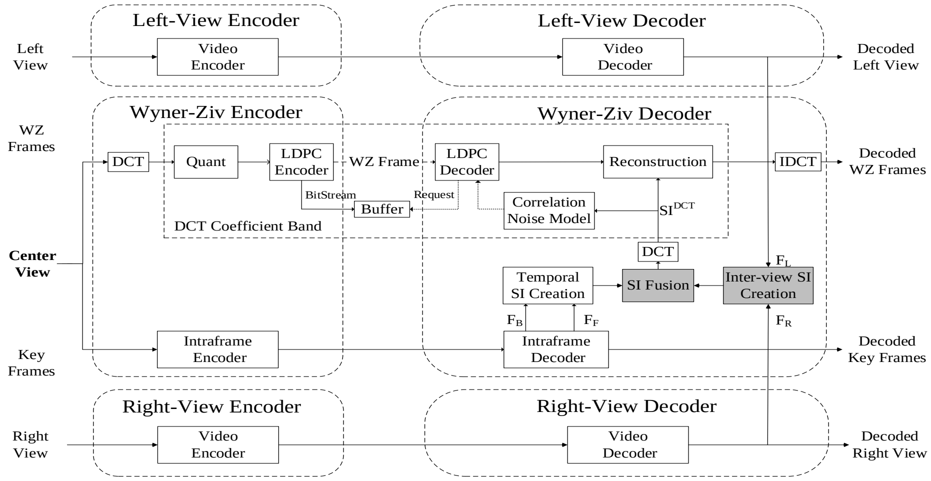


Fig.1. Framework of multi-view WZ codec.

3. Proposed Scheme

As illustrated in Fig.2, the proposed inter-view SI generation scheme for MDVC will be described in this section. The scheme includes five parts: luminance compensation, left/right disparity search, bidirectional disparity search, spatial disparity field smooth and compensation.

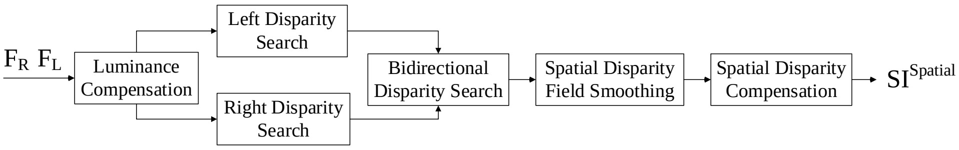

F_R F_L
Luminance Compensation
Left Disparity Search
Right Disparity Search
Bidirectional Disparity Search
Spatial Disparity Field Smoothing
Spatial Disparity Compensation
SI^Spatial

Fig.2. Block diagram of inter-view SI generation.

A. Luminance Compensation

In order to reduce the influence of illumination for view disparity search, luminance compensation for video frames at the left and right views is required. According to [14], by using the cumulative histogram of a distorted frame H_d and the cumulative histogram of the reference frame H_r, a mapping function can be defined as

$$M[i]=j,\ if\ H_r[j]\le H_d[i]\le H_r[j+1], \tag{1}$$

where i and j are bins of the cumulative histogram belong to H_d and H_r, respectively. Let $I_d(x,y)$ stand for a pixel at *(x, y)* location in a distorted frame. Then the compensated frame I_c can be derived by

$$I_c(x,y)=M\left[I_d(x,y)\right]. \tag{2}$$

For example, when luminance compensation is performed for the current frame L_t in the left view, we use the forward frame X_F in the central view as the reference frame, and the forward frame L_f in the left view as the distorted frame. Then the final compensated frame can be obtained by using Eq.1 and 2.

B. Left/Right View Disparity Search

The main purpose of left/right view disparity search is to estimate the view disparity vectors (DVs) between the left/right and the middle view. As illustrated in fig.3, based on 8×8 block level, we want to obtain the view disparity between the left and middle view for the current frame *T*. Let EV_B^{LM} denotes the displacement vector between the frame *T* in the left view and the frame *T-1* in the middle view, and EV_F^{LM} represents the displacement vector between the frame *T* in the left view and the frame *T+1* in the middle view. In the left view, by using the frame *T-1* as the reference frame, we can achieve the backward motion vectors MV_{t-1} of the current frame *T* through motion prediction. Let DV_{t-1} be the DV between the left and middle view for the frame *T-1*. Here EV_B^{LM} can be regarded as

$$EV_B^{LM}=MV_{t-1}+DV_{t-1}. \tag{3}$$

Similarly, we can also define the displacement vector EV_F^{LM} as

$$EV_F^{LM}=MV_{t+1}+DV_{t+1}, \tag{4}$$

where MV_{t+1} and DV_{t+1} denote the forward motion vectors of the current frame *T* in the left view, and the DV between the left and middle view for the frame *T+1*, respectively. It is assumed that the DV of video objects varies in a linear manner. Hence, we can obtain

$$MV_{t-1}=-MV_{t+1}. \tag{5}$$

And we can also get

$$2DV_t=DV_{t-1}+DV_{t+1}. \tag{6}$$

Here DV_t represents the DV between the left and middle view for the current frame *T*. Consequently, it can be derived as by using Eq.3, 4, 5, and 6

$$DV_t=\frac{EV_B^{LM}+EV_F^{LM}}{2}. \tag{7}$$

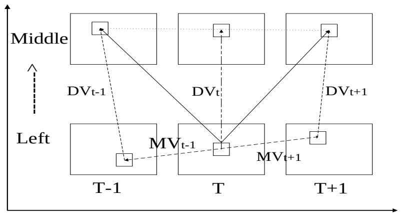


Fig. 3. Disparity mapping model

Next we will discuss how to obtain EV_B^{LM} and EV_F^{LM} accurately. We apply a hybrid search method that combines feature matching and epipolar line search. The details of the presented method are as follows.

Step 1: feature matching. We use the approach proposed in [16] to achieve feature points of blocks. Then the horizontal and vertical ranges for the epipolar line search are calculated.

Step 2: epipolar line search [17]. In this step, we firstly obtain the fundamental matrix by using the feature points founded in the Step 1. Then we search the epipolar line that belongs to the central point of matched blocks. Let *x* and *y* be the vertical and horizontal location of a pixel in the epipolar line, respectively. And the epipolar line can be calculated by

$$a * x + b * y + c = 0, \tag{8}$$

where *a*, *b*, and *c* are the corresponding parameters of the epipolar line. After that, using minimum sum of absolute difference (SAD) as measure, the candidate blocks can be determined by comparing the matched blocks obtained in step 1 with the ones in this step. Furthermore, some candidate blocks will be discarded when the distance between the candidate point and the epipolar line exceeds the horizontal and vertical range of the epipolar line search. And the distance can be calculated with

$$distance = \frac{|a * col + b * row + c|}{\sqrt{a^2 + b^2}}, \tag{9}$$

where *col* and *row* denote the vertical and horizontal coordinates of a candidate point, respectively.

Step 3: spatial smoothing. By using the displacement vectors of a candidate block and its neighboring ones as the input, the SAD value of each block for the reference and the matched frames would be computed. And when minimum SAD value is achieved, the corresponding displacement vector will be selected as best one for the current block.

C. Bidirectional Disparity Search

If we project the left/right view to the central view directly, many overlapped and uncovered regions will appear. To address this problem, we will change the reference view from the left/right to the central view. Besides the DVs obtained in Section B, we can also estimate the other two candidate DVs by matching the left and right view directly. Here we propose a multi-selection method to change the reference view and find out the optimal DVs for blocks. Details of the presented approach are as follows.

Step 1: initial selection. Since the locations of cameras are parallel, the video frames of the same time in different views should be also parallel. Thus, we can regard the candidate DVs whose horizontal coordinates are greater than a given threshold value as wrong vectors and discard them.

Step 2: range selection. In this step, the further search for candidate DVs is implemented by using the corresponding positions of vectors founded in Step 1 as the initial search locations. In the given search range, SAD values of each block in right and left views are calculated. For a block, when the SAD value is lower than σ, the corresponding vector would be selected as the candidate DV. And the corresponding SAD value is also recorded.

Step 3: optimal selection. For a block, we have obtained several candidate vectors through the search operations described in Section B and step 2. Then the DV whose SAD value is the minimum will be determined as the final DV. If the no candidate vector is achieved for a block, the final results will be set as a given value.

D. Spatial Disparity Field Smoothing and Compensation

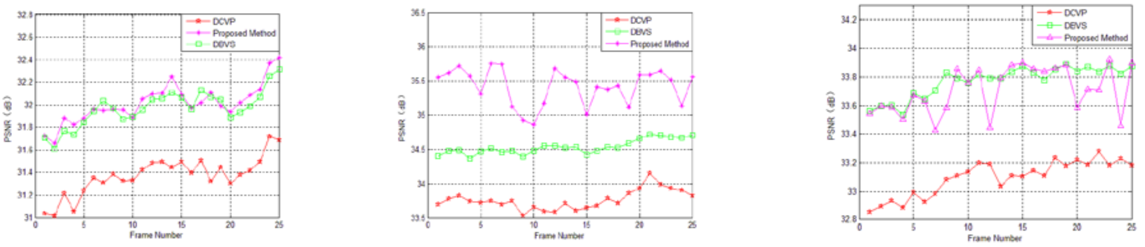
To improve the quality of spatial DVs, we apply the spatial disparity field smoothing (SDFS) algorithm [5] to smooth them. It is noted that we use the weighted median vector filter to deal with these DVs.

On the other hand, to identify and fix the existing disparity errors, spatial disparity compensation should be performed. Firstly, we find the cross-border blocks at the central view and replace them with the blocks which are at the same positions in the temporal SI frame. Then we further use the correspondences validation method [13] to achieve the unreliable blocks. Once the correspondence of a block is validated, the corresponding block in the central view will be replaced with the average pixel value of the two blocks in left and right view. In addition, for the unreliable blocks, we use the average value of the reliable DV candidates in the (8-connected) neighboring blocks as the DV of the corresponding blocks. If the surrounding DV candidates are invalid, the blocks will be replaced with the same blocks in the temporal SI. Thus, a spatial SI can be derived without camera intrinsic parameters.

4. Experimental Results and Discussion

To evaluate the performance of the proposed scheme, we conducted the experiments by employing the WZ codec for multi-view scenarios based on LDPC codes. Video sequences, whose resolutions are all 176×144, are utilized to the experiments. At the central view, the group of picture (GOP) and frame rate are set as 2 and 20 frames per second, respectively. Frames of the left and right views and key frames of the central view are coded using JPEG coding method. In addition, SURF algorithm [18] is applied to implement feature matching. The maximum distance of epipolar line search and σ are set 2 and 200, respectively. We use the peak signal noise ratio (PSNR) to measure the reconstructed video quality. In the experiments, we also implement DCVP [8] and DBVS [13] methods for comparisons.

Fig.4 shows the comparisons of decoded video quality for different video sequences under the same bitrates. From fig.4, we can observe that the reconstructed video quality of proposed scheme is superior to that of the DCVP method for all of video test sequences. Specially, the average PSNR gain obtained by the proposed scheme is up to 1.65 dB when testing the sequence exit. This reveals that the quality of inter-view SI generated by our scheme is better than that created by the DCVP method. And the reason is that the proposed scheme is capable of exploiting the spatial relation more efficiently than the DCVP method by utilizing temporal relation. On the other hand, it also can be seen that the decoded video quality of the presented scheme is roughly the same as that of DBVS approach from Fig.4. But unlike the DBVS approach, the presented scheme does not need to know the camera intrinsic parameters, thus it can be applied more flexible in practice.



(a) (b) (c)

Fig.4. PSNR comparisons for various video sequences. (a) ballroom. (b) exit. (c) vassar.

Fig.5 illustrates the second decoded frame of sequence vassar by using different methods. From Fig.5, we can see that the visual quality of reconstructed video employing our proposed scheme is better than that of the DCVP method obviously, while the same as that of the DBVS approach. This also indicates that our scheme can generate an inter-view SI frame for MDVC accurately.

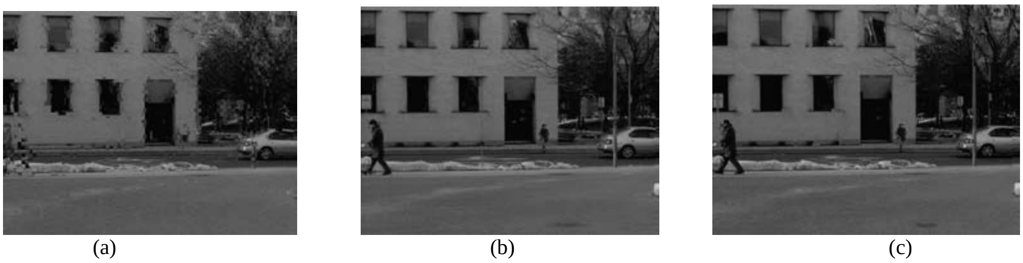


Fig.5. Visual quality of decoded video. (a) DCVP. (b) DBVS. (c) Proposed method.

5. Conclusion

To address the exiting issues among spatial SI creation techniques for MDVC, an inter-view SI generation scheme based on hybrid search and multi-selection is proposed in this paper. In the presented scheme, both feature matching and epipolar line searching are used to improve the accuracy of frame matching. Moreover, multi-selection has been used to obtain the optimal view DV. Experimental results show that the average PSNR value of decoded videos using the proposed scheme is higher than that of the DCVP method about 1.65dB at most, while 0.62dB at least for various test video sequences. On the other hand, without knowing the camera intrinsic parameters, the presented scheme is able to obtain the same decoded video quality compared with the DBVS approach.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61461006, No. 61261023); in part by the Guangxi Natural Science Foundation Project (No. 2013GXNSFBA019271); in part by the Scientific Research Fund of Guangxi Education Department (No. 201106LX016).

References

[1] D. Slepian, J. Wolf: IEEE Trans. Inf. Theory Vol.19(1973), p.471.
[2] A. Wyner, J. Ziv: IEEE Trans. Inf. Theory Vol.22(1976), p.1.
[3] B.Girod, A.M. Aaron, S. Rane, et al: Proc. IEEE Vol.93(2005), p.71.
[4] C. Brites, J. Ascenso, F.Pereira: Signal Process.: Image Commun. Vol.28(2013), p.689.
[5] J.Ascenso, C.Brites, F.Pereira: 5th EURASIP Conference on Speech and Image Process. (2005) p.1.
[6] J.Ascenso, C.Brites, F.Pereira: IEEE ICIP (2006), p.605.
[7] S.Ye, M. Ouaret, F.Dufaux, et al: IEEE ICIP (2008), p.2228.
[8] M.Ouaret, F.Dufaux, T.Ebrahimi:EUSIPCO (2007), p.3.
[9] C.Brites, F. Pereira: Signal Process.: Image Commun. (2015), p.1.
[10] H. Lv, H. Xiong, L.Song, et al: ICC (2009) p.1.
[11] H. Lv, H. Xiong, Y. Zhang Y, et al: IEEE ISCAS (2008), p.3450.
[12] X. Artigas, F.Tarrés, L.Torres:SIGMAP(2007), p.450.
[13] C. Brites, F.Pereira: IEEE Trans. Circuits Syst. Video Technol (2014), p. 1771.
[14] U.Fecker, M.Barkowsky,A.Kaup: IEEE Trans. Circuits Syst. Video Technol (2008) p.1258.
[15] D. G. Lowe: International journal of computer vision (2008), p.91.
[16] Information on http://blog.csdn.net/abcjennifer/article/details/7639681.
[17] R. Hartley, A. Zisserman: Cambridge Univ. Press (2000).
[18] B. Herbert, E. Andreas, T. Tinne, V. G. Luc: Computer Vision and Image Understanding Vol. 110 (2008), p. 346.