Distributed Multi-view Video Coding

Xun Guo[1, *], Yan Lu[2], Feng Wu[2], Wen Gao[1, 3], Shipeng Li[2]

[1]School of Computer Sciences, Harbin Institute of Technology, Harbin, 150001, China
[2]Microsoft Research Asia, Beijing, 100080, China
[3]Institute of Computing Technology, Chinese Science Academy, Beijing, 100080, China

ABSTRACT

There are mainly two key points which can affect the efficiency of multi-view video capture and transmission system largely: communication between cameras and computing complexity of encoder. In this paper, we propose a practical framework of distributed multi-view video coding, in which inter-camera communication is avoided and the large computing complexity is moved from encoder to decoder. In this scheme, multi-camera video sources are encoded separately and decoded dependently, and the traditional inter frame is replaced by Wyner-Ziv frame. To reach this goal, Wyner-Ziv theory on source coding with side information is employed as the basic coding principle. A Wyner-Ziv coding method based on wavelet transform and turbo codes is used as the core of the scheme. To further improve the coding performance, we also consider exploiting the large redundancy between adjacent views. A more flexible prediction method that can jointly use temporal and view correlations is proposed to generate the side information at the decoder. The experimental results show that the coding performance of proposed DMVC scheme is very promising compared to the traditional intra coding.

Keywords: Wyner-Ziv coding, multi-view video coding, wavelet transform, turbo codes

1. INTRODUCTION

It has been widely recognized that multi-view video coding (MVC) is one of the key technologies for a wide variety of future interactive multimedia applications, e.g. 3D television. Due to the large data volume, transmission the multi-view video requires much more bandwidth than traditional video. Consequently, how to efficiently compress multi-view video becomes more important than any other data. In the past decade, various MVC techniques have been developed. In [1], a multi-view matching cost and pure geometrical constraints algorithm is used to estimate disparity and to identify the occluded areas in the views. In [2], a sprite generation algorithm in multi-view sequences is proposed to improve coding efficiency. Recently, some techniques have been proposed to MPEG 3DAV group [3][4]. Since the multi-view video consists of video sequences captured by multiple cameras from different angles and locations, significant correlations may exist among views. Therefore, the common idea of the existing MVC techniques is to exploit the correlations between adjacent views in addition to the temporal and spatial correlations within a single view. In other words, inter-view prediction is usually performed during the encoding process. Although the inter-view prediction does improve the coding performance compared to the simulcast video coding, it still suffers from some shortcomings in the practical multi-view video capturing and transmission systems.

Firstly, the above inter-view prediction is based on the assumption that the video frames from different views can be freely exchanged or simultaneously available at the encoder. However, the communication between cameras with large data volume is normally very difficult in practice. Secondly, in a typical multi-view capturing system, all cameras are required to work simultaneously and video sequences are required to compressed and transmitted with low latency. Thus, the complexity becomes a big burden. As we know, most of the encoding complexity comes from the correlation exploration part, including the motion and/or disparity estimation. Is there any way to separately encode each frame of the multi-view video while the coding performance is still as good as jointly encoding? In theory, distributed source coding (DSC) can provide a solution to this problem. Theory of Slepian-Wolf shows that even if correlated sources are encoded without getting information from each other, coding performance can be as good as dependent encoding if the

[*] The work was done when the author was with Microsoft Research Asia as an intern.

compressed signals can be jointly decoded [5]. And also, Wyner and Ziv have extended the theory to the lossy source coding with side information [6]. Recently, several practical Slepian-Wolf and Wyner-Ziv coding techniques have been proposed for video coding, namely, distributed video coding (DVC). In [7], Pradhan and Ramchandran proposed a DVC framework based on syndrome of codeword coset. In [8], Aaron and Girod proposed a DVC scheme using turbo codes. In these schemes, temporal prediction is done when decoding instead of encoding and an asymmetric coding system is achieved.

In this paper, we further extend the DVC strategy to multi-view video, and propose a generic structure for distributed multi-view video coding (DMVC). In the proposed scheme, the multi-view video is taken as a 2D image matrix. Based on the pre-defined structure, each frame in the multi-view video can be independently encoded as either a traditional Intra-frame (I frame) or a Wyner-Ziv frame (W frame). Following the basic idea proposed in [8], the Wyner-Ziv frame is encoded using turbo codes and decoded with side information generated from the reference frames. A more flexible side information generation algorithm considering both temporal and view-directional correlations is proposed to achieve high prediction accuracy. In addition, wavelet transform is employed for the Wyner-Ziv frame coding for the purpose of, on the one hand, exploiting the spatial correlation, and on the other hand, utilizing the high-order statistical correlation. And because of the inherent level-structure of wavelet transform, the spatial and quality scalability can be achieved conveniently.

The rest of this paper is organized as follows. Section 2 presents the proposed distributed multi-view video coding framework including the whole structure, Wyner-Ziv coding part and a new generation strategy for side information. Experimental results are shown in Section 3 and conclusions are drawn in section 4.

2. PROPOSED DMVC SYSTEM

2.1 DMVC structure

Figure 1 shows the coding structure of the proposed DMVC system. In this system, multi-view video frames are classified into two categories: Intra frames and Wyner-Ziv frames, noted by I and W respectively. Intra frames are coded with the traditional DCT based intra coding method. Wyner-Ziv frames are inserted between successive two intra frames and the number can be adjusted according to the coding requirement. Because Wyner-Ziv frames can be intra encoded and inter decoded, the whole system consists of independent encoder and joint decoder. Thus, low encoding complexity and high coding performance can be achieved. Side information, noted by Y, plays an important role in this scheme, because it is the key part when decoding Wyner-Ziv frame. As shown in the figure, each Wyner-Ziv frame W needs a side information frame Y and Y is generated at the decoder through motion/disparity compensated prediction.

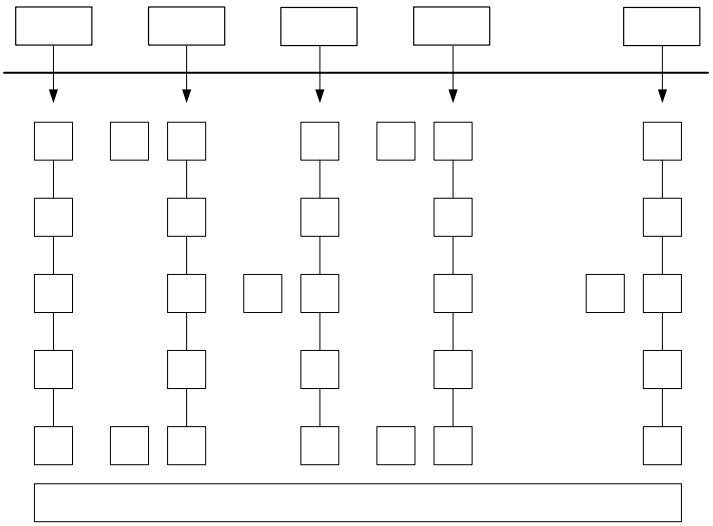


Figure 1: Coding structure of the proposed distributed multi-view video coding system.

The proposed DMVC system has mainly two advantages:

1) The communications between the different cameras can be removed. In the previous MVC schemes, inter-view correlations are exploited at the encoder through disparity compensated prediction method. However, in practical applications, this kind of data exchange between cameras is very difficult. Thus, inter-view prediction is almost impossible when using these schemes. But in proposed system, Wyner-Ziv frames are also independently encoded. Therefore, no communication has to be done between cameras. This advantage will be enlarged in the case of dense multi-camera system.
2) The number of views that need to be decoded is more flexible. In traditional MVC schemes based on hybrid video coding, if inter-view prediction is used, the number of reference frames is pre-decided when encoding. Thus, all the reference frames have to be decoded before the current frame no matter which view they are from. However, sometimes only one view is needed and the inter-view correlation is not so strong. In this case, the reference views still have to be transmitted and decoded for correctly decoding current view. The proposed DMVC scheme can avoid this redundancy, because the inter-view prediction is done at decoder and decoding adjacent views or not can be chosen freely.

2.2 Wyner-Ziv coding scheme

The basic theory for Wyner-Ziv coding is as follows. Let X and Y be statistically dependent signals, and let Y be known as side information for encoding X. Then the conditional rate-mean squared error distortion function for X is the same whether the side information Y is available only at the decoder or both at the encoder and the decoder. We try to utilize this characteristic in multi-view video coding scenario. Our goal is to achieve the best performance with the least complexity at encoder. Thus, we employ the basic idea proposed in [6], ie. source coding with side information using turbo codes.

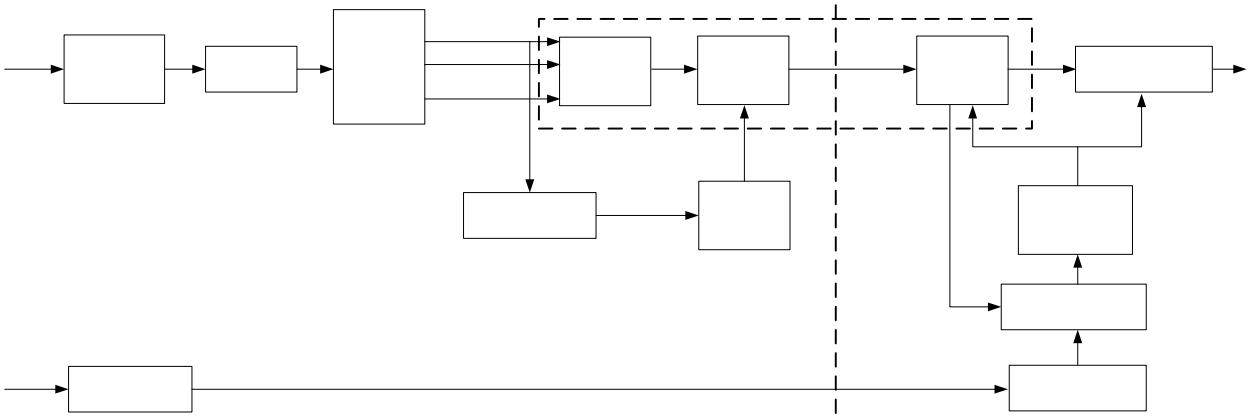
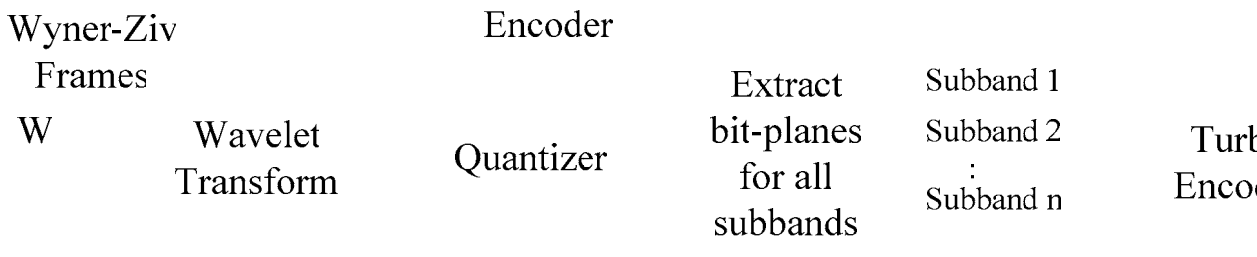


Figure 2: Block diagram of the encoder/decoder of the proposed distributed multi-view video coding.

Figure 2 shows the encoder/decoder diagram of the Wyner-Ziv coding scheme used in proposed DMVC scheme. Intra frames are encoded and decoded with 263+ intra coding scheme. Wyner-Ziv frames are coded using turbo codes based scheme included in the dashing block. The core of the Wyner-Ziv coder is a rate-compatible punctured turbo code (RCPT) [9]. The turbo code consists of two identical constituent convolutional codes and the generator matrix $[\frac{1+D+D^3+D^4}{1+D^3+D^4}]$ is used to generate parity bits. Wavelet transform is used on Wyner-Ziv frames to exploit the spatial correlation within a frame and improve the coding performance. We can describe the encoding and decoding process of Wyner-Ziv frames as follows.

Encoder

1) Wyner-Ziv frame W is decomposed into several levels by wavelet transform. In this paper, we use 9/7 wavelet filter and 3 decomposition levels. The decomposition structure is as shown in Figure 3.



2) The transform coefficients are quantized into M levels using a uniform scalar quantizer. Note that different subbands may have different M values. M is related to the selection of quantization step size for each subband. Because of the inherent hierarchical characteristic of wavelet transform, subbands in higher level, e.g. LL_3, give more contribution to reconstructed quality. Thus, we use the quantization ratio for all the subbands as the following equation:

$$q_i = \frac{Q}{\sqrt{\omega_i}}, \tag{1}$$

where Q represents the quantization parameter and i represents the subband number. ω_i is a weight value that represent the contribution of subband i, which can be computed from the wavelet filter.

3) The quantized coefficients are input into the turbo encoder bit-plane by bit-plane. Coefficients of the same subband are encoded together. Only the parity bits are stored in the parity buffer and transmitted to the decoder. A puncture matrix is employed to select parity bits.

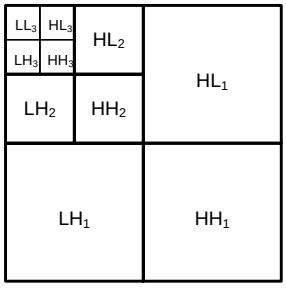


Figure 3: Wavelet decomposition structure used in Wyner-Ziv coding.

The rate allocation is always a tough task for Wyner-Ziv coding. In the original Wyner-Ziv coding scheme based on RCPT, the number of requested bits is usually decided through a feedback channel. In this paper, we employ a puncture schedule at the encoder instead of getting the feedback from the decoder. Particularly, the energy of the higher subbands is used to estimate the bit rate. Since the side information frame is usually generated from the global warping from adjacent views, the matching accuracy is most likely decided by the current frame complexity. Our current scheme sometimes allocates more bits so as to make sure the correct reconstructions. The more accurate rate allocation scheme remains our future work.

Decoder

At decoder, intra frames are decoded before the Wyner-Ziv frames between them. Side information is generated by using interpolation among the surrounding intra frames, which will be described in details in next subsection. After applying wavelet transform on side information Y, the received parity bits and the generated side information are used to decode and reconstruct the Wyner-Ziv frame. The decoder will successively decode the coefficients of a subband until an acceptable probability of bit error rate is achieved.

The turbo decoder and the reconstruction modules assume a Laplacian residual distribution between Wyner-Ziv frame W and side information Y. Let d be the difference between corresponding coefficients in X and Y. Then, the distribution of d can be approximated as $f(d) = \frac{\alpha}{2} e^{-\alpha|d|}$ for each subbands [8]. Let, c_j denotes the ith bit of a coefficient c_j, and $c_j^{i'}$ denotes the estimated reconstruction value for c_j^i. The probability P can be computed using the residual distribution model as follows:

$$P = \frac{\alpha}{2} e^{-\alpha \left| d_{C_j^i} \right|}, \text{ with } d_{C_j^i} = (m_i I(c_j^{i'}) + offset) - I(y_j). \tag{2}$$

Here m_i represents the magnitude of ith bitplane. $I(c_j^{i'})$ indicates the possible value of c_j^i, which is equal to 1 or 0. y_j is the coefficient of side information corresponding to c_j. *offset* is an estimated value used to compensate the lower part of c_j, because the lower bitpane of c_j is still not decoded now. The value of *offset* is decided according to the distribution parameter and the quantization step size. The currently decoded bitplane will be used to help decoding the next bitplane.

To further improve the coding efficiency, a re-estimation process is considered. Particularly, if the subbands in higher level are decoded correctly, they can be used to help the motion estimation and get a more accurate prediction as finer side information. And the finer side information can be used to better decode the remaining subbands.

2.3 Side information generation

Temporal direction

There are several methods to generate the temporal side information: directly using previously reconstructed frame, extrapolation method and interpolation method. In order to get the best performance, we use motion compensated prediction based interpolation method. Particularly, the prediction of the current Wyner-Ziv frame is generated from the forward and backward intra frames. This method is similar to the symmetric method in the prediction of traditional B frames. We assume that most of the motions in three successive frames are linear and the current motion vectors can be derived from the motion between adjacent two intra frames. Thus, motion compensation can be finished even if the current frame is absent.

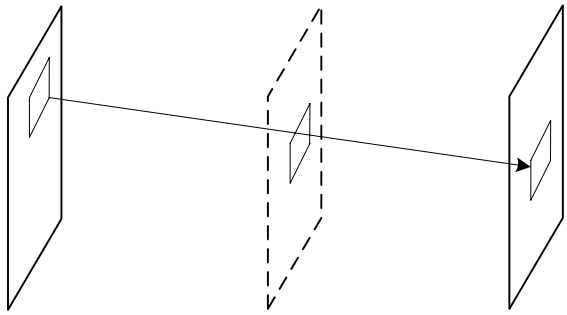


Figure 4: Motion-compensated interpolation for side information.

Figure 4 gives an example of the motion-compensated interpolation. We assume that I_{n-1} and I_{n+1} are intra frames coded with intra coding scheme. W_n is the Wyner-Ziv frame between the two intra frames. Now, we attempt to generate side information for W_n by using I_{n-1} and I_{n+1}. Note that we can not get information from W_n, thus we can only use the motion between I_{n-1} and I_{n+1} to estimate the motion of W_n. MV_{n-1} is the motion vector of a block in I_{n-1}. We assume the motion in these three frames are linear and the block position in W_n can be derived from MV_{n-1}.

View direction

In order to fully utilize inter-view correlations, we propose a prediction method for frames from adjacent views. Due to the special characteristic of multi-view video, frames at the same time instant in different view sequences are usually captured by cameras from different angles and locations. The global disparities among these frames can be represented by global motion models, which have been extensively used for pixel prediction in existing multi-view video coding schemes. In this paper, we use a six-parameter affine model which can be described as the following equation:

$$\begin{aligned} x' &= a_{11}x + a_{12}y + b_x \\ y' &= a_{21}x + a_{22}y + b_y \end{aligned}, \tag{3}$$

where x, y and x', y' represent the locations of current frame and reference frame respectively. a_{11}, a_{12}, a_{21}, a_{22}, b_x and b_y represent global motion parameters.

Mixed prediction

In the traditional MVC techniques based on the hybrid video coding, since the temporal correlation is usually stronger than the inter-view correlation, the motion-compensated prediction with motion vectors is more frequently used rather than the inter-view prediction. However, in a typical DMVC scenario, the motion vectors indicating the temporal

I_{n-1}

correlations do not exist and the predictive side information frame can only be interpolated from the adjacent intra frames, whereas the global camera parameters can usually be achieved. In other words, in the DMVC scheme, the inter-view correlation should be more helpful than the temporal correlation. Based on the proposed coding structure, at the decoder side, a Wyner-Ziv frame typically can have four reference frames, i.e. two from the adjacent views and another two from the same view.

Since the coding performance greatly depends on the accuracy of side information, we propose a more flexible side information generation algorithm, which can predict the current Wyner-Ziv frame from both temporal and view directions.

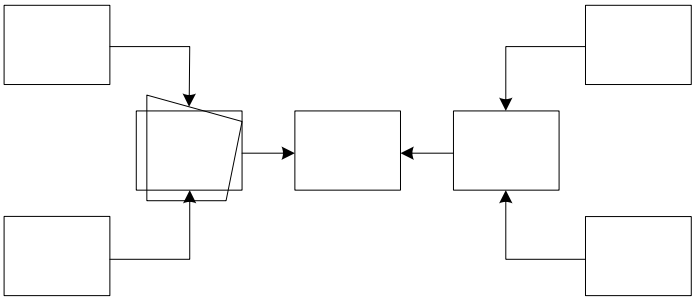


Figure 5: Temporal and View-directional prediction for Wyner-Ziv frames.

As shown in Figure 5, I_{n-1} and I_{n+1} represent the temporal adjacent frames, while I_l and I_r are the co-located frames from the left and right views. Let Y_1 and Y_2 denote the temporal prediction and the inter-view prediction, respectively. In temporal direction, the forward and/or backward reconstructed frames within the current view are used as reference frames and the interpolation method based on motion-compensated prediction is used to construct the predicting frame Y_1. In view direction, the co-located reconstructed frames in the adjacent views are used as the reference frames and global warping is used to construct the predicting frame Y_2. Obviously, the co-located regions in Y_1 and in Y_2 usually have the different prediction accuracy or reliability. Thus, the data fusion algorithm is employed to construct the more accurate predicting frame $Y = C(Y_1, Y_2)$. Particularly, the temporal interpolation is used only when the difference between the forward and backward predictions is smaller than a threshold T_1 and the estimated motion is smaller than a threshold T_2.

3. EXPERIMENTAL RESULTS

In order to verify the coding efficiency of the proposed scheme, experiments on real multi-view video sequences are carried out. Test sequence Race1 and Crowd provided by KDDI Lab are used in the test. In each sequence, three views with 128 frames (320x240) of each view are selected. In the test, we use the IWIW structure. Thus, every Wyner-Ziv frame can be predicted from its forward and backward Intra frames and another two Intra frames from adjacent views. The symmetric motion estimation method is used to compensate and interpolate the temporal side information. The 6-parameter global warping model is used to generate the view side information. Statistics show that the proposed joint temporal and inter-view prediction method can significantly reduce the prediction errors compared to simply using the temporal prediction. Figure 6 shows the side information frames generated from the temporal prediction and from the inter-view prediction, respectively. Obviously, the inter-view prediction works well for the high motion objects, e.g. the region in the white block; and the temporal prediction works well for the static region, e.g. the logo region. Although sometimes the side information also has the good visual quality, the bits of Wyner-Ziv frame target at removing the large local errors. Figure 6(c) shows a reconstructed Wyner-Ziv frame.

The absolute coding performance is also evaluated. Figure 7 shows the rate-distortion (R-D) curves of both sequences. Only the PSNR of luminance component and the bitrate of Wynzer-Ziv frames are illustrated. The curve of “263+ I frames” indicates the results of 263+ intra coding; the curve of “MC_Temporal” indicates Wynzer-Ziv coding with temporal prediction only; and the curve of “MC_GME” indicates the Wynzer-Ziv coding with the joint temporal and inter-view prediction. According to the figure, the proposed DMVC scheme with joint temporal and inter-view prediction outperforms H.263+ intra coding about 4 dB for Race1, and about 7 dB for Crowd. It also outperforms the DMVC with only temporal prediction about 1.5 dB in PSNR. Moreover, the PSNRs of the Intra frames for side

information generation are about 36 dB. According to the R-D curves, at the point of PSNR equivalent to 36 dB, more than 50% bits can be saved using the proposed algorithm compared to the H.263 intra+ coding.

4. CONCLUSIONS

In this paper, we have presented a novel distributed multi-view video coding scheme, in which the Wyner-Ziv theory on source coding with side information is employed as the basic coding principle. The wavelet transform and turbo codes are jointly utilized in the Wyner-Ziv frame coding. Moreover, a more flexible prediction method is proposed to favorate side information generation at the decoder. With the proposed DMVC scheme, the inter-camera communication is avoided and the large computing complexity is moved from encoder to decoder. Meanwhile, the coding performance is very promising compared to the traditional intra coding.

(a) (b)

(c)

Figure 6: (a) Side information generated from temporal interpolation; (b) Side information generated from inter-view prediction; and (c) Reconstructed Wyner-Ziv frame.

REFERENCES

1. R. S. Wang, Y. Wang, "Multiview Video Sequence Analysis, Compression, and Virtual Viewpoint Synthesis", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.10, pp.397-410, April 2000.
2. N. Grammalidis, M. G. Strintzis, "Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multiview Image Sequences", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.8. pp.328-344, June 1998.
3. H. Kimata and M. Kitahara, "Multi-view video coding based on scalable video coding for free-viewpoint video," ISO/IEC JTC1/SC29/WG11 M11571, Hong Kong, Jan. 2005.
4. Y. Ho, S. Yoon and S. Kim, "A framework for multi-view video coding using layered depth image," ISO/IEC JTC1/SC29/WG11 M11582, Hong Kong, Jan. 2005.

5. D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp.471-480, July 1973.
6. A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transaction on Information Theory*, vol. 22, pp.1-10, Jan.1976.
7. R. Puri, and K. Ramchandran, "PRISM: a new robust video coding architecture based on distributed compression principles," *Proc. of 40th Allerton Conference on Communication, Control, and Computing*, Allerton, IL, Oct. 2002.
8. A. Aaron, S. Rane and B. Girod, "Transform-domain Wyner-Ziv codec for video," *Proc. Visual Communications and Image Processing, VCIP-2004*, San Jose, CA, January 2004.
9. D. Rowitch and L. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo codes," *IEEE Transactions on Communications*, vol.48, no.6, pp.948-959, June 2000.

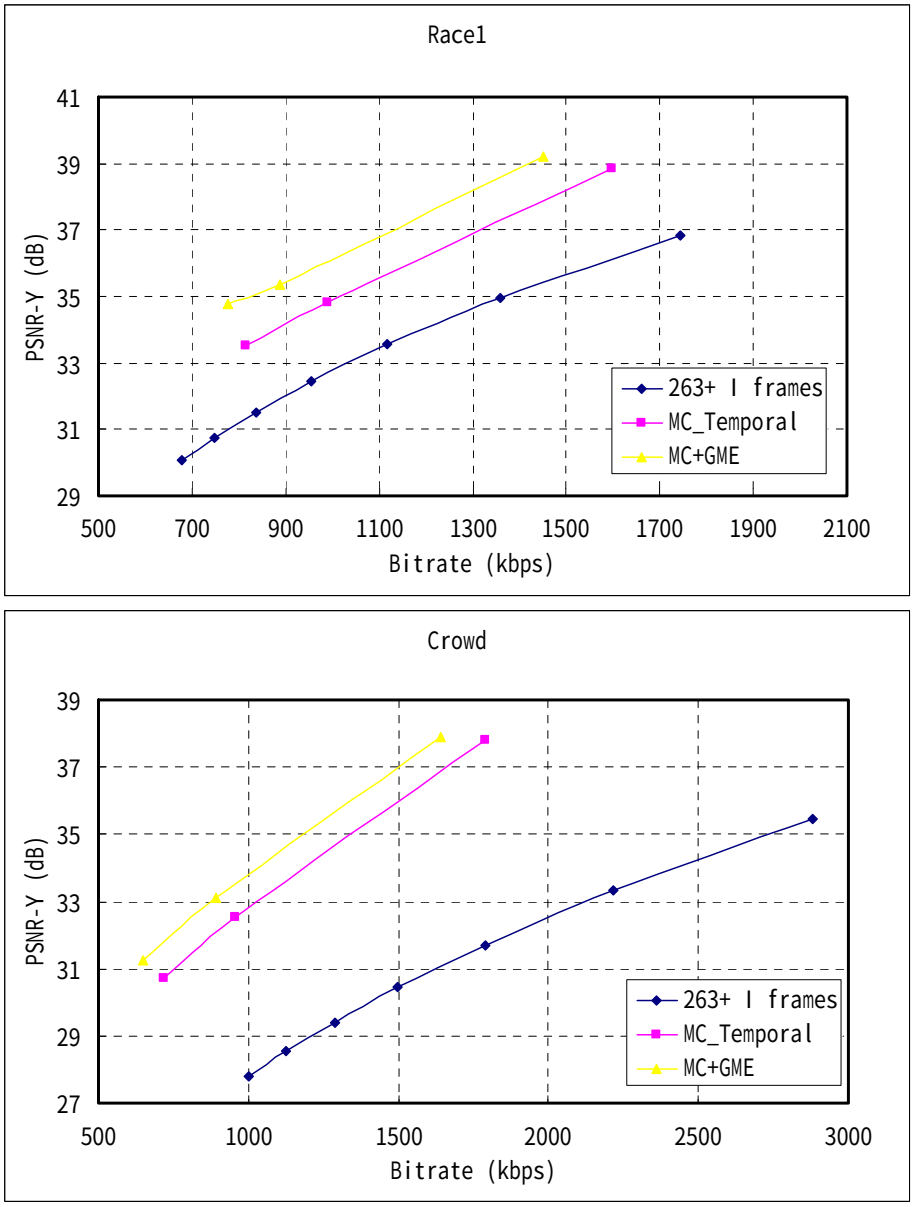


Figure 7: R-D curves of the Race1 and Crowd.