Region-based Fusion Strategy for Side Information Generation in DMVC

Yongpeng Li *[a], Xiangyang Ji[b], Debin Zhao[c], Wen Gao[d]
[a]Graduate University, Chinese Academy of Science, Beijing, 100039, China;
[b]Institute of Computing Technology, Chinese Academy of Science, Beijing, 100080,China
[c]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001,China
[d]Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, 100871,China
{ypli,xyji,dezhao,wgao@jdl.ac.cn}

ABSTRACT

The coding efficiency of Wyner-Ziv frames relies considerably on the quality of side information and the capability to model the statistical dependency between the original frame and side information. In the field of distributed multi-view video coding (DMVC), there are two kinds of side information, namely temporal side information (TSI), which is generated by exploiting the temporal correlation, and inter-view side information (IVSI), which is generated by exploiting the inter-view correlation. This paper proposes a new fusion method to get better side information by the region-based combination of TSI and IVSI. Besides, an improved statistical model of "correlation channel" is proposed to estimate the statistical dependency between the original frame and side information at the decoder. We call it Region-based Correlation Channel Model (RCCM). The RCCM models the "correlation channel" between original frame and side information at a fine granularity level by detecting the spatial quality variation of side information within each Wyner-Ziv frame. Experimental results demonstrate that the RCCM can more accurately model the "correlation channel" between the original frame and side information and thus, less bits are required for decoding a Wyner-Ziv frame.

Keywords: Distributed video coding, multi-view video coding, side information, correlation

1. INTRODUCTION

Distributed video coding, based on Slepian Wolf's [1] and Wyner and Ziv's [2] information theoretic results from the 1970s, is a new paradigm for video coding. In contrast to the convectional video coding paradigm, the inter-frame dependence is only exploited at the decoder in distributed video coding. Because the bulk of computation is shifted to the decoder, it is more suitable for applications like distributed sensor networks, cell-phone, etc, where the encoder has a limited amount of resource available. Recently, DVC is becoming increasingly popular, and a lot of practical distributed video coding solutions have been investigated. In [3], Pradhan and Ramchadram proposed a DVC framework based on syndrome of codeword coset. In [4], Aaron and Girod proposed a DVC scheme based on turbo codec.

Distributed multi-view video coding (DMVC) is the extension of DVC st[1]rategy to multi-view video and aims to efficiently code several views, which are captured by multiple cameras at different locations, by exploiting all kinds of redundancies at decoder. It provides an appealing solution for the applications like sensor networks and video surveillance, etc, in which several correlated view sequences have to be encoded and the resources like battery-power and computation-capability for the encoder are limited. Many works in terms of distributed multi-view video coding have been investigated. In [5], a wavelet-domain Wyner-Ziv coding scheme is applied on multi-view video. In [6], several fusion-based side information generation methods are compared.

The major difference between the mono-view distributed video coding and the multi-view distributed video coding is that there are two kinds of side information for Wyner-Ziv frames in DMVC, namely temporal side information (TSI)

1 This work was supported by National Science Foundation of China (60672088, 60736043)

and inter-view side information (IVSI), respectively. In this paper, temporal side information is generated by the motion-compensation-based temporal interpolation and inter-view side information is generated by exploiting the inter-view correlation in term of motion correlation of two adjacent views. Based on the overall analysis of TSI and IVSI, the Wyner-Ziv frame can be partitioned into two regions: the region with low motions in which side information is often well interpolated and the region with high and irregular motions in which side information is usually badly interpolated. According to the data partition, a new approach is proposed to fuse the TSI and IVSI. In this approach, two different methods are used to yield the fused side information for the two kinds of regions, the region with low motions and the region with high and irregular motions, based on the comparison of TSI and IVSI. Since DVC implies a statistical mind set, the coding efficiency of Wyner-Ziv coding solutions depends critically on the capability to model the statistical dependency between the original frame at the encoder and side information computed at the decoder. In previous works, a Laplacian distribution is usually used to model the statistical correlation between the original frame and side information at frame level. The parameter of the Laplacian distribution is updated frame by frame at the decoder, because the side information quality varies along the sequence. As mentioned above, the side information in the region with low motions within the frame is often well generated; while the side information in the region with high and irregular motions within the frame is usually badly generated. That is to say, the side information quality is not stationary even within a frame. However, this kind of variation is not considered in previous work. The variation of side information quality generally indicates the statistical variation of the “correlation channel”. Based on this idea, a new statistical model of “correlation channel” between the original frame and side information is proposed. We call it Region-based correlation channel model (RCCM). RCCM respectively estimates individual relevant parameter for the two kinds of regions generated by the data partition mentioned above through the correlation between current frame and previous frame. Because the RCCM can more accurately model the “correlation channel” between the original frame and side information, less bits are required for decoding a Wyner-Ziv frame.

The rest of paper is organized as follows. Section 2 describes the proposed strategy in detail. It includes region-based fusion strategy for side information generation and RCCM. Experiment results are shown in Section 3 and conclusions are given in Section 4.

2. THE PROPOSED STRATEGY

2.1 DMVC coding structure

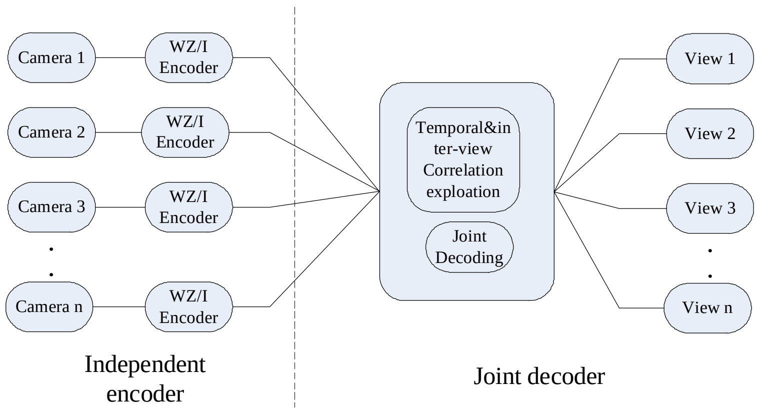

Fig. 1 Framework for DMVC

Fig. 1 illustrates the framework for DMVC. In this framework, each view is encoded separately without communicating with each other. In each view sequence, each frame is encoded either as a key frame or as a Wyner-Ziv frame. Key frames are coded using traditional DCT-based intra coding method. The pixel-domain Wyner-Ziv coding scheme described in [4] is adopted for Wyner-Ziv frames. The core of the scheme is a rate-compatible punctured turbo codec (RCPR), which consists of two identical constituent convolution codec and uses a generator matrix to generate parity bits.

As shown in Fig.2, for a Wyner-Ziv frame, each pixel is first uniformly quantized with 2^M intervals. Then the quantized pixels are input into the Slepian-Wolf coder, which is implemented using a Rate-Compatible Punctured Turbo code. At the decoder side, side information is generated by exploiting temporal correlation or inter-view correlation between current frame and the decoded key frames and, possibly decoded Wyner-Ziv frames. After that, the generated side information in conjunction with the received parity bits are used to reconstruct the original frame. If the decoder cannot reliable decode the original stream, it requires additional parity bits from the encoder through feedback until an acceptably probability of bit error is guaranteed.

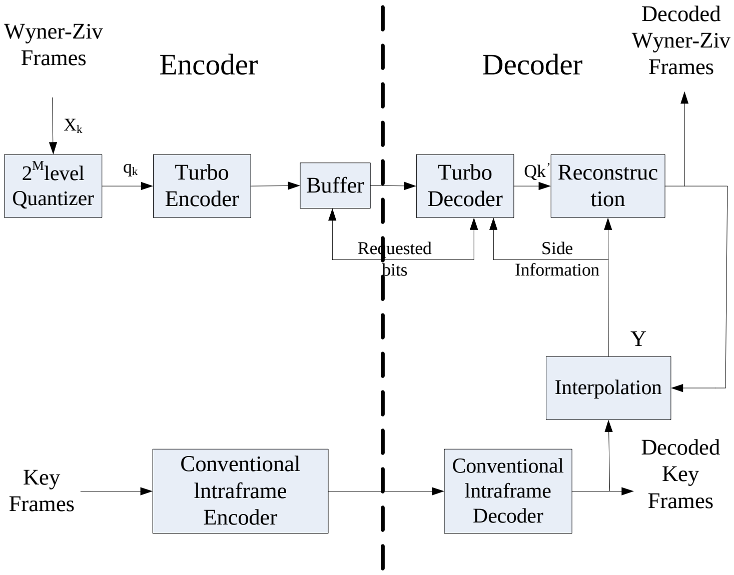

Fig. 2 Pixel-domain Wyner-Ziv codec

2.2 Fusion based side information generation

2.2.1. TSI generation

Temporal interpolation (TI) is the general side information generation method in DVC schemes. TI generates the side information by exploiting the temporal correlation that exists between the temporally successive frames under the translational model. This method is similar to the symmetric method in the prediction of traditional B frames. As shown in Fig. 3, W_n is a Wyner-Ziv frame in time n; I_{n-1} and I_{n+1} are key frames that are temporally adjacent to W_n. The target of temporal interpolation is to generate side information for W_n by the motion-compensated prediction although it is absent. Temporal interpolation works under the translational model, in which we assume that most of the motions in three successive frames are translational. For forward prediction, the motion vector of block b_i, denoted by MV_f, can be derived from the motion vector of the co-located block in I_{n-1} with the equation $MV_f = \frac{1}{2} MV_{n-1}$. Similarly, we can get backward prediction motion vector $MV_b = \frac{1}{2} MV_{n+1}$. Then the prediction value of b_i, denoted by P, can be yielded through the equation $P_b = \frac{1}{2}(P_{MV_f} + P_{MV_b})$

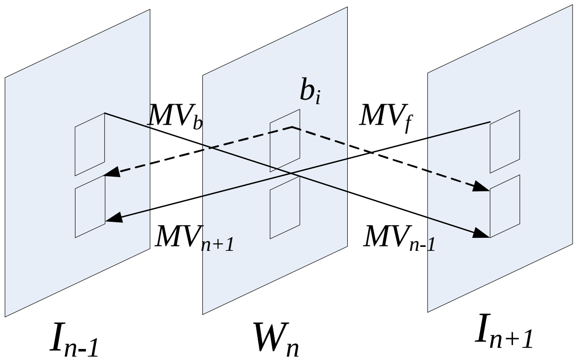

Fig. 3 Temporal interpolation

2.2.2. IVSI generation

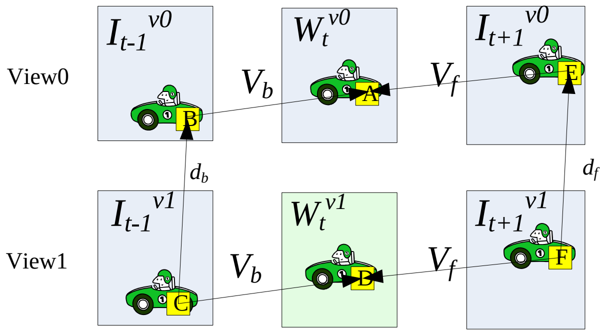

Fig. 4 Motion correlation

In this paper, the inter-view side information (IVSI) is generated by exploiting the inter-view correlation in term of motion correlation of two adjacent views. Fig.4 is an illustration of motion correlation between two adjacent inter-view frames. The form of F_t^v indicates coding information of a frame. Specially, F denotes the coding scheme used for the frame: I indicates that the frame is a key frame and is intra coded; while W indicates that the frame is Wyner-Ziv coded. The superscript v expresses the view index (view 0 or view 1) and the subscript t specifies the time instant of the frame. For example, I_{t-1}^{v0} indicates that the frame in view 0 captured at time instant t-1 is intra coded. In the scenario, $W_t^{v0}, I_{t-1}^{v0}, I_{t-1}^{v1}, I_{t+1}^{v0}$ and I_{t+1}^{v1} are reconstructed and the target is to generate side information for W_t^{v1}. The generation of IVSI is based on motion correlation of two adjacent views. The basic idea for motion correlation is that the same object should have similar motion trend in two adjacent inter-view frames if the two views are captured by two parallel cameras. More specially, the basic idea for motion correlation is that the motion vector of corresponding block in adjacent view is used as an estimate of the motion vector of current block. As shown in Fig. 4, view 0 and view 1 are two view sequences captured by two parallel cameras. C is a block in I_{t-1}^{v1}. It can be seen that, block B, the corresponding block of C in I_{t-1}^{v0}, can be determined by the disparity d_b which is obtained by performing disparity estimation on I_{t-1}^{v0} and I_{t-1}^{v1}. As we know, the block B's motion trend towards W_t^{v0}, denoted by V_b, can be obtained by performing motion estimation on W_t^{v0} and I_{t-1}^{v0}. As mentioned above, the motion vector of block B can be used as an estimate motion vector of block C. Here, V_b is used as an estimate motion vector for block C. By V_b, the block D in W_t^{v1} can be determined , for which block C is used as an prediction in term of pixel value. In other words, the side information for the blocks in W_t^{v1} can be thus generated by this progress. We called this progress backwards motion derivation, because side information for W_t^{v1} is obtained by its backwards frame I_{t-1}^{v1}. And corresponding side information is called backwards derivation side information (BDSI). Similarly, forwards derivation side information (FDSI), which is generated for W_t^{v1} through W_t^{v0}, I_{t+1}^{v0} and I_{t+1}^{v1}, can be further generated to improve the side information quality. Accordingly, the side information for W_t^{v1} is determined as follows:

$$SI_t^{v1} = \begin{cases} (SI_{t\_forwards}{}^{v1} + SI_{t\_backwards}{}^{v1} + 1)/2 & (a) \\ SI_{t\_forwards}{}^{v1} & (b) \\ SI_{t\_backwards}{}^{v1} & (c) \\ SI_{t\_TI}{}^{v1} & (d) \end{cases} \tag{1}$$

where SI_t^{v1} denotes the finally generated IVSI for b, which is a block in W_t^{v1}. $SI_{t\_forwards}{}^{v1}$ and $SI_{t\_backwards}{}^{v1}$ express FDSI for b and BDSI for b, respectively. As shown in (1), if both BDSI for b and FDSI for b can be derived, the finally generated side information for b is yielded by averaging them (corresponding to (a) in (1)), otherwise the only obtained BDSI or FDSI is set as the final side information. Because neither BDSI nor FDSI can be obtained for the block located in the area in the frame that is non-overlapped with the adjacent views, the side information for such blocks is yielded by $SI_{t\_TI}{}^{v1}$, which denotes the side information generated by using TI. Therefore, the side information for each block in the W_t^{v1} can be generated, and the target of generating side information for W_t^{v1} is thus achieved.

A more detailed derivation structure for IVSI generation is shown in Fig. 5, in which the sequences including three views are coded, the three views are: view 0, view 1, and view 2, respectively. The coding structure IWIW is adopted for each view. Since view 1 is the middle view, which is overlapped with both view 0 and view 2, we apply the proposed region-based fusion strategy and RCCM on the Wyner-Ziv frames in view 1. In this case, it can be seen that two kinds of IVSI can be yielded for the Wyner-Ziv frames in view 1, they are $SI_{through\_view0}^{v1}$ and $SI_{through\_view2}^{v1}$. $SI_{through\_view0}^{v1}$ is generated by exploiting inter-view correlation between view 1 and view 0; while $SI_{through\_view2}^{v1}$ is generated by exploiting the inter-view correlation between view 1 and view 2. Here, the final IVSI for the Wyner-Ziv frames in view 1 is improved by simply averaging $SI_{through\_view0}^{v1}$ and $SI_{through\_view2}^{v1}$ as follows:

$$SI_{IVSI}^{v1} = (SI^{v1}{}_{through\_view0} + SI^{v1}{}_{through\_view2} + 1)/2 \tag{2}$$

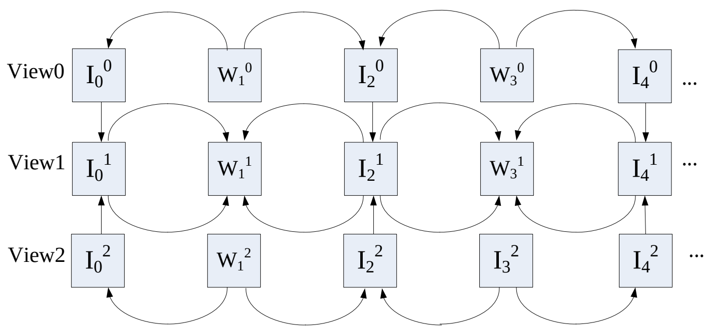

Fig. 5 Derivation for IVSI generation

2.2.3. Fusion approach

Temporal interpolation usually performs well in the region with low motions within the frame, such as static background, because the translational model assumed in TI is probably satisfied in such region. TSI in this region is usually well generated and the residuals between TSI and original frame are small in this region. However, in the region where the translational model is not satisfied, TSI usually performs poorly. This region usually corresponds to the areas with high and irregular motions. In such region, IVSI should perform much better in comparison with TSI. This is because IVSI can obtain more accurate motions from the adjacent decoded inter-view frames than TSI. Fig. 6 is an illustration of comparison of TSI and IVSI in the region where translational model is not satisfied. As shown in the figure, the block for which the side information is to be generated in current frame is denoted by b. The motion vectors derived by using TI for b from forward prediction and backward prediction (as described in section 2.2.1) are denoted by MV_{TSI_f} and MV_{TSI_b}, respectively. As the translational model assumed in TI is probably not satisfied, MV_{TSI_f} and MV_{TSI_b} are probably not accurate; while in IVSI, side information for b is generated by exploiting the inter-view correlation in term of motion correlation that is described in section 2.2.2. That is to say, the motion of bock b is estimated by that of its corresponding block c in view 0. Note that c is decoded and reconstructed by requiring a subset of parity bits from the encoder. This means that the error pixels in c are corrected largely. Hence, MV_{IVSI_f} and MV_{IVSI_b}, which are obtained by performing motion estimation on block c and its adjacent forward and backward frame in view 0, are more closed to the actual ones. Based on motion correlation model of two adjacent views described in section 2.2.2, more accurate motion vectors for block b can be derived accordingly. So IVSI is more closed to the original frame compared to TSI in such region.

Based on the analysis above, it can be seen that both TSI and IVSI works well in the region with low motions where translational model is satisfied; while IVSI works better than TSI in the region with high and irregular motions where the translational model is probably not satisfied. Hence, a method that can exactly identify the region with low motions and the region with high and irregular motions within the frame is very helpful to the efficient fusion of TSI and IVSI. This is hard problem because the original frame is not accessible at the decoder side. In this paper, we propose a method to identify the region with low motions and the region with high motions within the frame based on the overall analysis and comparison of TSI and IVSI. Specially, the residual between TSI and IVSI is used as the criterion to distinguish the region with low motions from region with high and irregular motions. Because the both TSI and IVSI work well in the region with low motions, the region in which the residuals between TSI and IVSI are small is identified as the regions with low motions. The rest is identified as the region with high motions. The detailed principle for identify the two kind of regions is described as follows:

$$\begin{cases} if \sum_{\vec{v}\in S} \left| SI_{temporal}(\vec{v}) - SI_{\text{int}erview}(\vec{v}) \right| \le T, S \in R_0 \\ else, S \in R_1 \end{cases} \quad (3)$$

where S is comprised of all sampling position of the macroblock in current frame and T is a predefined threshold. $SI_{temporal}$ and $SI_{\text{int}erview}$ stand for TSI and IVSI, respectively.

Since both TSI and IVSI are well generated in R_0, the fused side information in R_0 is yielded by averaging $SI_{\text{int}erview}$ and $SI_{temporal}$ for the purposed of de-nosing; while TSI is bad in R_1, and IVSI performs much better than TSI in R_1, the fused side information in R_1 is yielded by IVSI. The details of the strategy are described as follows:

$$\begin{cases} SI(\vec{v})_{fused} = (SI_{temporal}(\vec{v}) + SI_{\text{int}erview}(\vec{v}) + 1)/2, \vec{v} \in R_0 \\ SI(\vec{v})_{fused} = SI_{\text{int}erview}(\vec{v}), \vec{v} \in R_1 \end{cases} \quad (4)$$

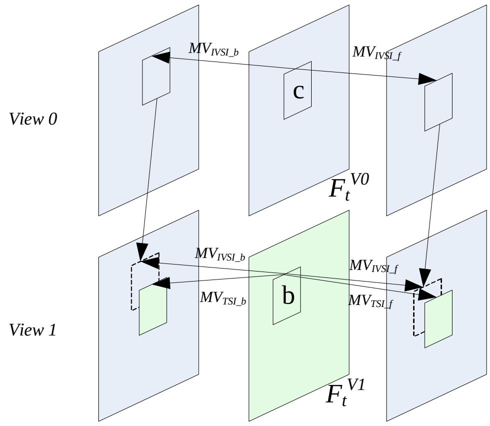

Fig. 6 Comparison of TSI and IVSI

2.3 RCCM model

Another thing that the coding efficiency of Wyner-Ziv frame significantly depends on is the capability to model the statistical dependency between the original frame and side information. The statistical dependency in Wyner-Ziv coding is modeled by a virtual "correlation channel" with an error pattern characterized by some statistical distribution (or model) e.g. Laplacian distribution as shown in (5):

$$f(WZ - SI) = \frac{\alpha}{2} e^{\alpha |WZ - SI|} \tag{5}$$

where WZ stands for the original frame and SI expresses side information.

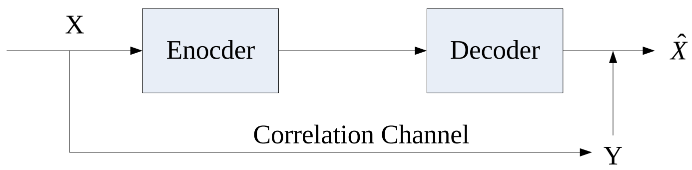

Fig. 7 Correlation channel

Fig. 7 is the illustration of the "correlation channel". The side information Y calculated at the decoder is viewed as a "corrupted" version of the original frame X, and the decoding progress targets at correcting the errors in the "corrupted" frame. If the "correlation channel" accurately describes the relationship between WZ-SI, the coding efficiency is high. Otherwise, if the "true" correlation is significantly different from the estimate one, coding performance loss will be observed.

Generally speaking, good interpolation of side information means that the residuals between the original frame and side information are small. While, bad interpolation of side information means that the residuals between the original frame and side information are high. As we know, the side information quality varies along the sequence, and such variation of the side information quality can be viewed as the result of the statistical variation of the "correlation channel" between the original frame and side information. In previous works, to correspond to such statistical variation of the "correlation channel", the parameter of the Laplacian distribution is updated frame by frame at the decoder.

However, the statistical correlation between the side information and the original frame are not spatially stationary within a frame. Usually, the side information generation algorithms often failed in the regions with high motion objects and the errors in side information for such region increase significantly; but, it is quite good in the regions with low motion, such as static background. So the regions where the interpolation is successful and the regions where the interpolation has failed coexist within the side information frame. And this also can be viewed as the result of the statistical variation of the “correlation channel” between side information and original frame, which is not considered in previous works.

In this paper, an improved statistical model of “correlation channel” is proposed to describe “correlation channel” at a fine granularity level. We call it Region-Based correlation channel model (RCCM). In the RCCM, different “correlation channel” are assumed in different the regions according to the side information quality in the region. Specially, in the R_0, described in section 2.2.3, since the side information is closed to the original frame, the “correlation channel” is characterized by a Laplacian distribution with a narrow α_0. While, in R_1 where the residuals between the side information and original frame are high, the “correlation channel” is characterized by a Laplacian distribution with wide α_1. Both α_0 and α_1 for current frame are estimated by observing the statistics from the latest decoded Wyner-Ziv frame at the decoder side. Here, we take the estimation of α_0 as an example to describe the progress. First, the variance of the residuals between the latest reconstructed Wyner-Ziv frame and its side information in R_0 is calculated as follows:

$$\sigma_0{}^2 = E\left[(R(\vec{v}) - SI(\vec{v}))^2\right] - (E\left[(R(\vec{v}) - SI(\vec{v}))\right])^2, \vec{v} \in R_0 \tag{6}$$

where R and SI indicate the latest reconstructed Wyner-Ziv frame and its side information, respectively. $\vec{v}$ denotes the sample position in R_0. $E[.]$ is the expectation operator.

After that, α_0, which is the parameter of Laplacian distribution in R_0, is calculated by (7)

$$\alpha_0{}^2 = \frac{2}{\sigma_0{}^2} \tag{7}$$

The same progress is conducted to calculated α_1 .Then in the progress of decoding the current Wyner-Ziv frame, the condition probability as shown in (8), is used.

$$\begin{cases} p(x/y) = \dfrac{\alpha_0}{2} e^{\alpha_0|y-x|}, \vec{v} \in R_0 \\ p(x/y) = \dfrac{\alpha_1}{2} e^{\alpha_1|y-x|}, \vec{v} \in R_1 \end{cases} \tag{8}$$

where $\vec{v}$ denotes the sample position in the frame; α_0 and α_1 is the estimated parameter of Laplacian distribution for R_0 and R_1, respectively.

3. EXPERIMENT RUSULTS

In this section, we evaluate the rate distortion performances of the region-based fusion strategy and the RCCM. Experiments are carried out on multi-view sequences *Balllroom* and *Race1*, each of which is composed of 3 views in resolution of 640x480 and each view contains 60 frames. Both *Balllroom* and *Race1* are captured by the cameras that are spaced in parallel manner. The threshold T in the date partition progress (described in section 2.2.2) is set to 8. In the test, we use IWIW structure for each view and apply the region-based strategy and RCCM on view 1 (the central view). The details derivation structure of IVSI generation for the Wyner-Ziv frames in view 1 is illustrated in Fig.5.

Fig.8 shows the rate-distortion (RD) curves of both sequences, in which only the PSNR of luminance component and bitrates of Wyner-Ziv frame in the central view (view 1) are illustrated. The curve of “Framelevel_TSI” represents that side information is temporally generated and the statistical dependency between side information and the original frame is modeled by traditional Lapalican distribution at frame level. For the further comparison, the curves of “Framelevel_FusedSI” and “RCCM_FusedSI” are also demonstrated. These two methods represent that the dependency between side information and the original frame is modeled by traditional Lapalican distribution and the RCCM, respectively when the side information is generated using the fusion approach. It can be observed that the

“Framelevel_FusedSI” can obviously outperform the “Framelevel_TSI”. “RCCM_FusedSI” is able to further improve coding efficiency compared to “Framelevel_FusedSI”.

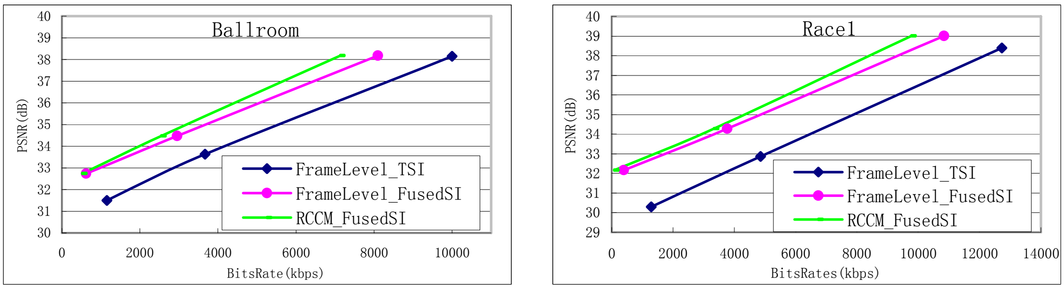

Fig. 8 R-D curves for *Ballroom and Race1*

4. CONCLUSIONS

In this paper, a new fusion strategy to fuse TSI and IVSI is proposed for side information generation in DMVC. In the proposed strategy, the residual between TSI and IVSI are used as the criterion to identify the regions with low motions and the region with high and irregular motions. The fused side information for the region with lows is yield by averaging TSI and IVSI for the purpose of de-noising; while the fused side information in the region with high and irregular motion is yielded by IVSI. Furthermore, a new statistical correlation model, called region-based correlation channel model (RCCM), is proposed. In RCCM, the “correlation channel” between the original frame and side information is modeled at a fine granularity level by detecting the spatial quality variation of side information within each Wyner-Ziv frame. Specially, the “correlation channel” in the region where side information is well interpolated is modeled by a narrow Laplacian distribution; while the “correlation channel” in the region where side information is not well interpolated is modeled with a wide Laplacian distribution. Experimental results demonstrate that the RCCM can more accurately model the “correlation channel” between the original frame and side information and thus, less bits are required for decoding a Wyner-Ziv frame.

REFERENCES

1. J.D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources”, *IEEE Transaction on Information Theory*, vol. *IT*-22, pp.471-480, July. 1973.
2. A.D. Wyner and J. Ziv, “The rate-distorition function for source coding with side information at the decoder”, *IEEE Transaction on Information Theory*, vol. 22, pp.1-10, Jan. 1976.
3. R. Puri, and K. Ramchandran, “PRISM: a new robust video coding architecture based on distributed compression principles”, *Proc .of 40th Allerton Cofference on Communication, Control, and Computing*, Allerton, IL, Oct. 2002.
4. A. Aaron, Zhang and B. Girod, “Wyner-Ziv coding for Motion video”, Asilomar Coference on Signals, Systems and Computers, Pacific Grove, USA, November 2002.
5. X. Guo, “Distributed Multi-View Video Coding”, *Proc. of SPIE, VCIP 2006*, San Jose, California, USA, Vol. 6077, pp. 290-297, January 2006.
6. Xavi Artigas, Egon Angeli, and Luis Torres, “Side information generation for Multiview distributed video coding using a fusion approach”, 7th Nordic Signal Processing Symposium, June 7-9, 2006, Reykjavik, Iceland.