Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Side information generation with auto regressive model for low-delay distributed video coding

Yongbing Zhang [a,b,*], Debin Zhao [a], Hongbin Liu [a], Yongpeng Li [c], Siwei Ma [d], Wen Gao [d]

[a] *Harbin Institute of Technology, Harbin, China*
[b] *Graduate School at Shenzhen, Tsinghua University, Beijing, China*
[c] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[d] *Peking University, Beijing, China*

ARTICLE INFO

Article history:
Received 24 March 2009
Accepted 3 October 2011
Available online 14 October 2011

Keywords:
Distributed video coding
Low delay
Auto regressive model
Side information
Wyner-Ziv
Probability based fusion
Centrosymmetric rearrangement
Least mean square

ABSTRACT

In this paper, we propose an auto regressive (AR) model to generate the high quality side information (SI) for Wyner–Ziv (WZ) frames in low-delay distributed video coding, where the future frames are not used for generating SI. In the proposed AR model, the SI of each pixel within the current WZ frame t is generated as a linear weighted summation of the pixels within a window in the previous reconstructed WZ/Key frame $t-1$ along the motion trajectory. To obtain accurate SI, the AR model is used in both temporal directions in the reconstructed WZ/Key frames $t-1$ and $t-2$, and then the regression results are fused with traditional extrapolation result based on a probability model. In each temporal direction, a weighting coefficient set is computed by the least mean square method for each block in the current WZ frame t. In particular, due to the unavailability of future frames in low-delay distributed video coding, a centrosymmetric rearrangement is proposed for pixel generation in the backward direction. Various experimental results demonstrate that the proposed model is able to achieve a higher performance compared to the existing SI generation methods.

1. Introduction

With the development of high performance computing and channel coding [1], distributed video coding (DVC) has received more and more attentions in recent years due to its desirable properties for some applications such as wireless low power video surveillance, video compression and sensor networks. DVC is based on the principles stated by Slepian–Wolf [2] for the lossless case and Wyner–Ziv (WZ) [3] for the lossy scenario. The majority of Slepian–Wolf and WZ coding systems adopt channel coding principles [4–7], assuming the statistical dependence between the two correlated sources X and Y as a virtual binary symmetric channel or additive white Gaussian noise channel. Compression of the source X can be achieved by transmitting only parity bits using error correcting codes. At the decoder side, with the aid of received parity bits and Y, called the side information (SI) of X, the error correcting decoding is performed, i.e., performing MAP or MMSE estimation of X.

Based on these theorems, some practical DVC systems have been presented. Pradhan and Ramchandran proposed a constructive and practical framework for distributed source coding using syndromes (DISCUS) [4] to perform WZ coding. Puri and Ramchandran proposed a power-efficient, robust, high-compression, syndrome-based multimedia (PRISM) [8] DVC framework. Besides, Aaron et al. provided an asymmetric WZ coding scheme [9] for motion video using intra-frame encoding and inter-frame decoding. In their framework, the key frames are encoded by H.263+ intra frame mode and the WZ frames are encoded by Slepian–Wolf codec based on turbo codes.

One of the most critical aspects in enhancing the compression efficiency of DVC is improving SI quality. According to the Slepian–Wolf theorem [2], the less the conditional entropy $H(X|Y)$ is, the fewer the bits to reconstruct X are required, under the condition that Y can be perfectly reconstructed at the decoder. Intuitively, in practical system, where SI is generated at the decoder side, better SI will result in better performance for the WZ frames. Different from the most existing video compression standards, where the computationally intensive motion estimation is performed at the encoder side, DVC shifts the motion estimation to the decoder side. Consequently, it is very difficult to generate high quality SI without the existence of the original video sequence at the decoder side.

According to the way SI generated, DVC can be categorized into interpolation and extrapolation cases. In interpolation case, similar to the B frame coding in hybrid video coding, SI is generated by the interpolating between the previous and following reconstructed

[*] Corresponding author at: Graduate School at Shenzhen, Tsinghua University, Beijing, China.

E-mail addresses: ybzhang@jdl.ac.cn (Y. Zhang), dbzhao@jdl.ac.cn (D. Zhao), hbliu@jdl.ac.cn (H. Liu), ypli@jdl.ac.cn (Y. Li), swma@jdl.ac.cn (S. Ma), wgao@jdl.ac.cn (W. Gao).

1047-3203/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2011.10.001

WZ/key frames [10–15]. On the contrary, in the extrapolation case, the SI is generated by referring only the previous reconstructed frame [16–22]. Generally speaking, the SI generated by interpolating has superior performance than that generated by extrapolating, since the former can use the future information to generate SI. However, this only holds if the temporal distance is small enough [20], i.e. the GOP (group of pictures) size is sufficiently small. Besides, the extrapolation DVC is very desirable in the sequential decoding for low latency cases, since the decoding process begins as soon as it receives the previous reconstructed frame, without waiting for the arrival of the following reconstructed key frame.

To improve the compression performance of low-delay DVC, many pioneering works have been done to improve the quality of SI. In Natario's scheme [19], a robust extrapolation module is proposed to generate SI based on motion field smoothening. In this method, the extrapolation is completed by motion estimation, motion field smoothening, motion projection as well as overlapping and uncovered areas. Borchert et al. [20] introduced a true motion based extrapolation scheme considering the 3-D recursive search (3DRS) motion estimation. All these methods resort to conventional motion estimation to extract motion information from the reconstructed video frames at the decoder side. They are all based on a translational motion model, in which it is assumed that the motion in the current frame is a continuous extension of the motion in the previous frame. However, the translation model is not always satisfied, especially for the video sequences with high motion.

To obtain higher quality SI in low delay DVC, in this paper we propose an auto regressive (AR) model based SI generation based on our previous work [22]. In the proposed AR model, the SI of each pixel within the current WZ frame t is generated as a linear weighted summation of pixels within a window in the previous reconstructed WZ/K frame $t-1$ along the motion trajectory. To capture the variation properties of the current WZ frame, the SI is generated block by block. The motion trajectory of each block is assumed to be that of the co-located block in the previous reconstructed frame and is of integer-pixel accuracy. In order to obtain accurate SI, we use the forward derivation and backward derivation to compute two weighting coefficient sets for each block within the current WZ frame t. In the forward derivation, each reconstructed pixel within the collocated block in WZ/K frame $t-1$ is approximated as a linear weighted summation of pixels within the corresponding window in the reconstructed WZ/K frame $t-2$. The Least-Mean-Square (LMS) algorithm is then employed to derive the first coefficient set of the AR model. In the backward derivation, each pixel in the reconstructed frame $t-2$ can be approximated as the weighted summation of corresponding pixels in the reconstructed frame $t-1$. By the centrosymmetric relation of the backward and forward derivations, the second coefficient set is derived. Finally, a probability based fusion is proposed in which the SI of the processing block within the current WZ frame t is generated as the fusion of the two regression results, generated by using the two derived coefficient sets, as well as the traditional extrapolation result. It should be noted that the proposed AR model employs the pixels centered around the pixel indicated by the motion trajectory to perform extrapolation rather then the pixels centered around the collocated pixel as in [23,24]. In addition to, the proposed AR model exploits the centrosymmetric property of the AR model to further improve the extrapolation accuracy. To verify the superiority of the proposed AR model based SI generation for the low-delay DVC, various experiments are conducted and the simulation results have confirmed that the proposed method is able to achieve SI with much higher accuracy compared with other existing methods.

The reminder of this paper is as follows. The overall architecture of the proposed system is first presented in Section 2. Then the model description and the forward and backward derivations are described in detail in Section 3. The probability based fusion is given in Section 4 followed by the experimental results and analysis in Section 5. Finally the conclusions are provided in the last section.

2. Framework overview

The block diagram of the proposed AR model based low-delay DVC is depicted in Fig. 1. The coding process starts by dividing the input frames into key frames and WZ frames. At the encoder side, the key frames are encoded using the H.264/AVC intra coding scheme. The WZ frames are encoded by applying the 4 × 4 H.264/AVC DCT transform and the DCT coefficients of the entire frame are grouped together in DCT bands. Each DCT band is uniformly quantized and the bit planes are sent to the turbo encoder. The turbo coding procedure for the DCT bands starts with the most significant bit planes and generates the respective parity bits which are stored in the buffer and transmitted in small amount upon decoder request.

At the decoder side, the key frames are decoded using H.264/AVC intra decoding scheme. For the WZ frames, the SI is first generated by the proposed AR model. As shown in Fig. 1, the SI generation is composed of three modules: traditional extrapolation and the interpolations by two AR coefficient sets. In the extrapolation, the motion of each block in the current WZ frame t is derived by performing motion estimation between the reconstructed frames $t-1$ and $t-2$. The first coefficient set of the AR model is computed by the forward derivation and the second coefficient set of the AR model is computed by the backward derivation followed by the centrosymmetric rearrangement. Both the first and second set coefficients are then used to generate the SI through interpolation process. Results of the three modules are then combined by a probability based fusion to generate the final SI. Then the iterative turbo decoder uses the received parity bits to correct the SI errors and generates the decoded quantized symbol stream. Finally, IDCT is applied to generate the WZ decoded frames.

3. Model description and its forward and backward derivations

In this section, we will first give the detail description of the proposed AR model, and then we will present the forward and backward derivations to compute two reliable AR coefficient sets so as to generate high quality SI.

3.1. Model description

In the proposed AR model, the SI of each pixel within the current WZ frame t is generated as a weighted summation of the pixels within a particular window in the previous reconstructed WZ/K frame $t-1$ as shown in Fig. 2. Let $\mathbf{X}_t$ be the current WZ frame at t, and $\mathbf{Y}_t$ be the SI of $\mathbf{X}_t$. For each pixel in $\mathbf{X}_t$, the window, indicated by the circles and the red arrow in Fig. 2, is determined by the integer-pixel accuracy motion field estimated during the motion extrapolation. After the determination of the window, the weighted summation is performed as

$$Y_t(m,n) = \sum_{-R \le (i,j) \le R} \hat{X}_{t-1}(\tilde{m}+i, \tilde{n}+j) \bullet \alpha(i,j). \tag{1}$$

Here $Y_t(m,n)$ represents the SI of the pixel located at (m,n), $\hat{\mathbf{X}}_{t-1}$ represents the previous reconstructed frame $t-1$, $(\tilde{m},\tilde{n})$ represents the corresponding integer-pixel position in $\hat{\mathbf{X}}_{t-1}$ determined by the motion vector of $\mathbf{X}_t(m,n)$, which is obtained during the motion extrapolation, $\alpha(i,j)$ is the forward AR coefficient from frame $t-1$ to frame t. In Eq. (1), R is defined to be the radius of the window, the size of which is $(2R+1)\times(2R+1)$. The proposed AR