Distributed Video Coding Based on the Human Visual System

Yongpeng Li, Debin Zhao, Siwei Ma, and Wen Gao, *Fellow, IEEE*

***Abstract*—This letter proposes a distributed video coding (DVC) scheme based on the human visual system (HVS). As we know, HVS can seldom sense any changes below the just-noticeable-difference (JND) distortion threshold due to its underlying sensitivity and masking property. Therefore, the un-noticeable signal differences between the original frame and side information need not to be corrected in DVC. Experimental results demonstrate that the proposed algorithm can save the bits-rates significantly without degrading the subjective quality of the reconstructed frames.**

***Index Terms*—Distributed video coding, human visual system.**

I. Introduction

In conventional hybrid video coding paradigm, high-complex motion compensation is performed at the encoder to achieve higher compression, while the decoding has relative low complexity. It generally suits a scenario such as broadcasting and video on demand, in which videos are encoded only once but are decoded several times. However, in some emerging applications like low power wireless video surveillance and multimedia sensor networks, the encoder has a limited amount of resources available and thus cannot employ complex motion compensation.

The requirements of these applications can be fulfilled with distributed video coding (DVC). DVC is a novel coding paradigm, in which the majority computational bulk can be shifted from the encoder to the decoder. The foundation of DVC is Slepian-Wolf (SW) theory [1] and the Wyner-Ziv (WZ) theory [2].Slepian and Wolf first proved that although two statistical sources X and Y are independently encoded, for lossless coding, similar coding performance can be achieved as long as the joint decoding of them is allowed. Wyner and Ziv extended the theory to the lossy case with side information at the decoder.

In DVC, side information is regarded as a corrupted version of the original frame, and the signal differences between the original frame and side information i.e., the errors in side information, are corrected by the parity bits sent from the encoder. In the literature, a variety of methods have been proposed to save parity bits by exploiting the spatial correlation within frame [4], [5] or improving side information quality [6]–[8]. Some schemes with block classification [10]–[12] are also proposed to improve rate distortion performance. In these works, the original frame was recovered by removing errors for the whole side information frame.

As we know, the human visual system (HVS), which is the ultimate receiver of the decompressed video signal, can seldom sense any changes below the just noticeable distortion (JND) threshold around a pixel due to their underlying temporal/spatial sensitivity and masking properties [3]. Therefore, it is unnecessary to spend parity bits on correcting the well-estimated signals. Based on this idea, this letter proposes a distributed video coding scheme based on HVS. The main novelty of the scheme is introducing JND model to DVC to save the un-necessary parity bits while retaining the visual quality of the reconstructed WZ frames. In the proposed scheme, a rough side information is generated at the encoder, which involves low-complexity computing, to examine which blocks of the original frame can be well estimated based on the JND model. Only the blocks that cannot be well estimated by the rough side information are encoded actually, that is to say, parity bits are generated only for these blocks. The remaining blocks will not be coded and will be recovered at decoder by their rough side information. Experimental results demonstrate that bit rates can be significantly saved without degrading the subjective quality of reconstructed WZ frames.

The rest of this letter is organized as follows. Section II presents the proposed scheme. The experimental results are shown in Section III, and the conclusion is given in Section IV.

II. Proposed Scheme

Fig. 1 illustrates the block diagram of the DVC scheme based on HVS. In the proposed scheme, each frame is encoded as either key frame or WZ frame. Key frame is coded using the H.264 Intra coding method. For the coding of WZ frame, the visual evaluation is performed referred as coding mode decision, in which the blocks within the WZ frame are classified into two categories according to their visual sensitivity (JND threshold): the actually-coded block and the copy block. The actually-coded blocks are the blocks that the difference between them and their referenced blocks in side information can be sensed. They are encoded using the pixel domain turbo-based WZ coding method. The copy blocks, on the other hand, are the blocks that the differences between them and their referenced blocks in side information cannot be perceived. They are not coded, and are recovered at the decoder by their side information directly. To help the decoder identify the locations of the actually-coded blocks and the copy blocks, the coding mode map is also entropy-coded and transmitted to the decoder. In

Manuscript received May 26, 2009; revised June 30, 2009. First published July 21, 2009; current version published August 26, 2009. This work was supported in part by National Science Foundation (60736043) and National Basic Research China (973 Program, 2009CB320905). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lisimachos Paul Kondi.

Y. Li is with the Graduate University, Chinese Academy of Sciences, Beijing 100085, China (e-mail: ypli@jdl.ac.cn).

D. Zhao is with Harbin Institute of Technology, Beijing 100085, China (e-mail: dbzhao@jdl.ac.cn).

S. Ma and W. Gao are with Peking University, Beijng 100085, China (e-mail:swma@jdl.ac.cn; wgao@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2009.2028111

1070-9908/$26.00 © 2009 IEEE

Authorized licensed use limited to: Peking University. Downloaded on November 23, 2009 at 05:14 from IEEE Xplore. Restrictions apply.

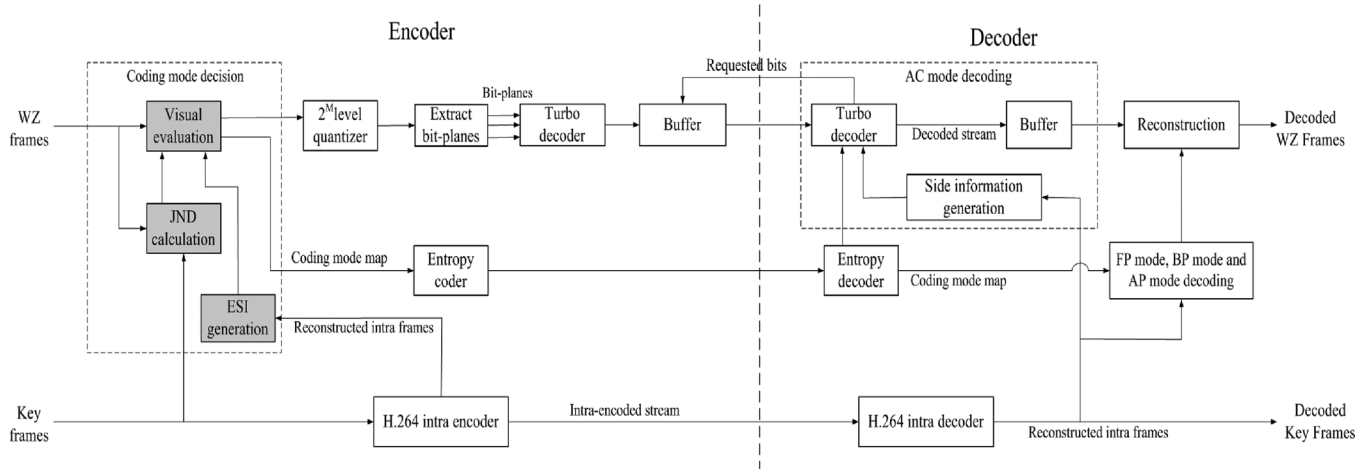

Fig. 1. Framework of the distributed video coding based on HVS.

the following, the coding mode decision, the encoding and decoding method of the actually-coded blocks, and the reconstruction method of the copy blocks will be described in detail, respectively.

Since successive frames in video sequence are strongly correlated with each other, certain parts of the current WZ frame can be estimated by its neighbor frames so well that the differences between the estimation, which is referred to side information in the scenario of DVC, and the original WZ frame can seldom be sensed by HVS. Intuitively, these well-estimated signals need not to be corrected. The JND model [9] is a practical measurement of the visibility of the signal differences. As the regions that can be well estimated within the current WZ frame are needed to be identified at the encoder, a rough side information termed encoder-side side information (ESI) is generated. For each block, three candidates, the co-located block in forward key frame, the co-located block in backward key frame, and the average of them, are firstly generated. The candidate that has the minimum sum of absolute difference (SAD) with the original WZ block is selected as the ESI of the block.

The classification of the actually-coded block and copy block is based on the comparison of the distortion of ESI and the spatio-temporal JND thresholds of the original WZ frame. The distortion of the ESI is calculated as

$$D(x,y) = |ESI(x,y) - I(x,y)| \tag{1}$$

where $D(x,y)$, $ESI(x,y)$, and $I(x,y)$ denote the distortion, the ESI intensity, and the original intensity of the pixel located at (x,y), respectively .

The spatio-temporal JND thresholds are calculated according to [9]. As stated, the spatial JND thresholds of image are determined by two factors, background luminance adaptation and texture masking. The effects of these two factors are integrated through the nonlinear additivity model as

$$JND^s(x,y) = T^l(x,y) + T^t(x,y) - C \times \min\{T^l(x,y), T^t(x,y)\} \tag{2}$$

in which JND^s refers to spatial JND thresholds, and $T^l(x,y)$ and $T^t(x,y)$ denote the visibility thresholds determined by the background luminance adaptation factor and the texture masking factor, respectively; (x,y) represent the pixel co-ordinates, and C is a constant in [0, 1], which plays an overlapping role. $T^l(x,y)$ reflects the characteristic that HVS is more sensitive to the luminance contrast rather than the absolute luminance, it can be calculated by

$$T^l(x,y) = \begin{cases} 17\left(1 - \sqrt{\frac{\bar{I}_Y(x,y)}{127}}\right) + 3, & \text{if } \bar{I}_Y(x,y) \le 127 \\ \frac{3}{128}(\bar{I}_Y(x,y) - 127) + 3, & \text{otherwise} \end{cases} \tag{3}$$

where $\bar{I}_Y(x,y)$ is the background luminance at (x,y). The texture related factor $T^t(x,y)$ represents the reflection of the characteristic that HVS is more sensitive to the errors in the smooth areas than those in texture areas, and it can be computed as

$$T^t(x,y) = 0.117 \times G(x,y) \times W(x,y) \tag{4}$$

where $G(x,y)$ denotes the maximal weighted average of the gradient around the pixel at (x,y), which can be computed by edge detection followed with Gaussian low-pass filter, and $W(x,y)$ is an edge-related weight of the pixel at (x,y), which can be computed by edge detection followed with Gaussian low-pass filter. In the scenario of video, the temporal affection on JND values should be taken into account. As a result, the spatio-temporal JND can be incorporated as the scaled amplitude of the spatial JND as shown in (5).

$$JND(x,y,t) = f(idl(x,y,t))JND^s(x,y) \tag{5}$$

where $idl(x,y,t)$ denotes the average inter-frame luminance difference between the tth frame and $(t-1)$th frame , and $f()$ is the empirical scaled amplitude function [9].

In the comparison, the blocks, which have more than 10% pixels whose distortion exceeds their corresponding spatio-temporal JND values, are identified as actually-coded block and are assigned with actual encoding (AC) mode. The remaining blocks are identified as copy block, and one of the three modes including forward prediction (FP) mode, the backward prediction (BP) mode, and the averaging prediction (AP) mode, which indicate the ESI is generated from the co-located block in forward key frame, the co-located block in backward key fame, and the average of them, respectively, is assigned to each copy block

Authorized licensed use limited to: Peking University. Downloaded on November 23, 2009 at 05:14 from IEEE Xplore. Restrictions apply.

Fig. 2. Reconstructed image of the 22nd WZ frame of *Paris* by the anchor technique.

Fig. 3. Reconstructed image of the 22nd WZ frame of *Paris* by the proposed technique.

to inform the decoder the generation method of its ESI. The decision threshold of 10% is empirically determined.

As aforementioned, the signal differences of the actually-coded blocks should be corrected. This is implemented by the pixel domain turbo-based WZ coding scheme presented in [4]. Pixels in these blocks are first re-organized together and quantized into 2^M levels using a uniform scalar quantizer. Then, the quantized pixels are fed into a turbo codec, which consists of two identical constituent convolution codecs. Finally, the parity bits generated by the turbo codec are stored in the buffer. At the decoder, the side information of the blocks with AC mode, which is generated by carrying out the motion compensated interpolation on the key frames decoded using the H.264 decoding algorithm, is fed into the turbo decoder. The turbo decoder is composed of two soft-input soft-output (SISO) decoders implemented using the maximum *a posteriori* (MAP) algorithm. Parity bits are successively requested by the turbo decoder through a feedback channel until a predefined threshold (10^{-3} in our proposed system) of bit-error rate is satisfied.

The copy blocks, which are assigned with FP mode, BP mode, and AP mode, are recovered by their co-located block in forward key frame, co-located block in backward key frame, and the average of the co-located blocks in backward and forward key frames, respectively.

III. Experimental Results

To evaluate the efficiency of the proposed algorithm, simulation experiments are carried out on four CIF sequences: *News, Paris, Tempete* and *Silent*. The anchor technique that we compare with is the pixel domain turbo-based WZ coding scheme presented in [4], in which all the blocks in WZ frame are WZ-encoded and side information of each WZ frame is generated using the motion compensated interpolation. The GOP size is 2 for both the proposed and anchor techniques, with even frame encoded as key frame and odd frame as WZ frame. The block size of 16×16 is adopted in the coding mode decision progress in the proposed technique.

We first examine the reconstructed images by the anchor technique and the proposed technique, which are shown in Figs. 2 and 3, respectively. For both images, the QP of the corresponding key frames adjacent of them is 28, and the quantizer of WZ coding is 8. As can be seen, it is hard to tell the differences between these two images, although the PSNR comparison of them is distinct: 36.32 dB for anchor technique versus 34.56 dB for the proposed technique, respectively. This indicates that the proposed algorithm avoid the penalty of subjective quality degradation. And notably, the coding of the image using anchor algorithm costs 69 708 bits, while the coding using the proposed algorithm costs only 21 598 bits. All of these are mainly attributed to the capability of visual sensitivity discrimination provided by the JND model.

To further test the efficiency of the proposed algorithm, a comparison in term of the bit-rates and the visual distortion is conducted, in which the subjective quality i.e., the visual distortion of the luminance component of WZ frame are compared for a variety of bit rate points. In the simulation, the QP of key frames for both techniques is 28, and to achieve similar bit rate range, the quantizers of the WZ frame for the proposed and the anchor techniques are $2^M \in \{2,4,8,16\}$ and $2^M \in \{2,4,8\}$, respectively. The subjective quality of the reconstructed frame is measured with the metric proposed in [9], i.e., the peak signal-to-perceptual ration (PSPNR) that only takes into account the distortion that exceeds the JND profile. The computation of PSPNR is shown as shown in (6), at the bottom of the next page, with

$$\delta(x,y,t) = \begin{cases} 1, & \text{if } |I(x,y,t) - \hat{I}(x,y,t) \geq JND(x,y,t)| \\ 0, & \text{other} \end{cases} \tag{7}$$

where $I(x,y,t)$ and $\hat{I}(x,y,t)$ denote the original and the reconstructed intensity of the pixel located at (x,y) in the tth frame, respectively. The experimental results for *News* and *Paris*, and *Tempete* and *Silent* are given in Figs. 4 and 5, respectively, in

Authorized licensed use limited to: Peking University. Downloaded on November 23, 2009 at 05:14 from IEEE Xplore. Restrictions apply.

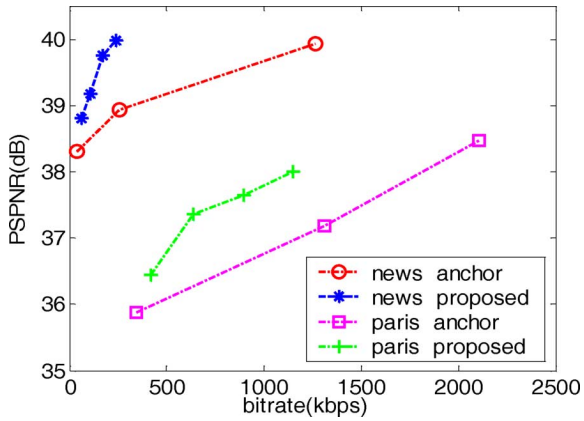

Fig. 4. Performance comparison of *News* and *Paris*.

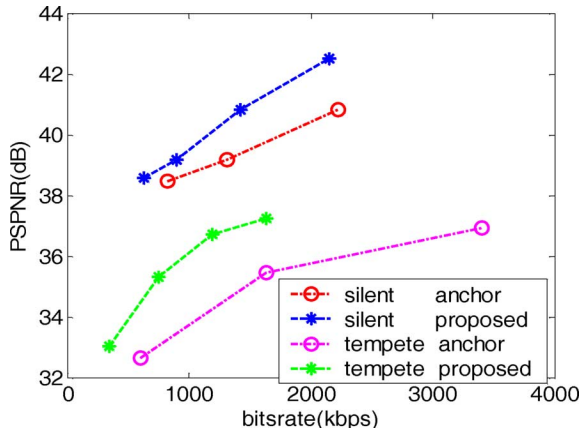

Fig. 5. Performance comparison of *silent* and *tempete*.

which the first 50 WZ frames in each sequence are evaluated. For all the sequences, clear reductions of bit-rates can be observed when the equivalent subjective quality is achieved. This substantially validate that the efficiency of the proposed algorithm in terms of the bit-rates and visual distortion performance.

The additional encoding complexity of the proposed scheme mainly comes from the generation of ESI and the calculation of JND values. On average, the generation of ESI involves additional encoding complexity by about 7 addition/subtractions and 1 division per pixel, and calculating JND value for a pixel needs about 48 addition/subtractions, 31 multiplications, 9 divisions, 2 square root operation and 1 arctangent operation. Compared with the motion compensation (ME) task performed in hybrid video coding, which requires about $(4 \times (n^2+1) \times R^2 + 4 \times (n^2+1) \times R + n^2) \times N_{\mathrm{mode}}/n^2$ additions/subtractions per pixel when one reference frame is employed, where R, n and N_{mode} denotes the search range, block size, and number of coding modes, respectively, the additional encoding complexity introduced by ESI generation and JND calculation is limited, roughly about 8%.of that of ME, when $R = 16$, $n = 16$ and $N_{\mathrm{mode}} = 4$, including P16 × 16, P16 × 8, P8 × 16 and P8 × 8 mode. This is also validated by the comparison of their actual execution time on the computer with Pentium 4 2.8-GHz CPU and 1 GB memory for CIF sequence: about 87 ms per frame for ESI generation and JND calculation, and about 1004 ms per frame for ME task, respectively.

IV. Conclusion

This letter proposes a distributed video coding based on the HVA, in which the JND model is introduced to identify the noticeable signal difference between the original frame and side information, and only noticeable signal differences are corrected. Experimental results show that the proposed scheme can significantly improve the WZ coding efficiency in terms of bit rate and subjective quality performance.

References

[1] J. D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 7, pp. 471–480, Jul. 1973.

[2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[3] N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[4] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[5] X. Guo, Y. Lv, F. Wu, and W. Gao, "Distributed video coding using wavelet," in *Proc. 2006 IEEE Int. Symp. Circuits and Systems*, Kos, Greece, May 2006, pp. 5427–5430.

[6] S. Klomp, Y. Vatis, and J. Ostermann, "Side information interpolation with sub-pel motion compensation for Wyner-Ziv decoder," in *Proc. Int. Conf. Signal Processing and Multimedia Applications (SIGMAP)*, Setúbal, Portugal, Aug. 2006.

[7] X. Artigas and L. Torres, "Iterative generation of motion-compensated side information for distributed video coding," in *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sep. 2005.

[8] J. Ascenso, C. Brites, and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding," in *Proc. 5th EURASIP Conf. Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic, Jun. 2005.

[9] X. K. Yang, W. S. Ling, Z. K. Lu, E. P. Ong, and S. S. Yao, "Just noticeable distortion model and its application in video coding," *Signal Process.: Image Commun.*, pp. 662–680, 2005.

[10] M. Tagliasacchi, A. Trapanese, S. Tubaro, J. Ascenso, C. Brites, and F. Pereira, "Intra mode decision based on spatio-temporal cues in pixel domain Wyner-Ziv video coding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

[11] Y. Wang, J. Jeong, and C. Wu, "Wyner-Ziv video coding with block classification," in *Proc. Int. Conf. Multimedia and Expo.*, Hannover, Germany, Jun. 2008.

[12] L. Liu, D. He, A. Jagmohan, A. Lu, and E. J. Delp, "A low-complexity iterative mode selection algorithm for Wyner-Ziv video compression," in *Proc. IEEE Int. Conf. Image Processing*, Palma de Mallorca, Spain, Oct. 2008.

$$PSPNR(t) = 10\log_{10}\frac{255 \times 255}{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(|I(x,y,t) - \hat{I}(x,y,t)| - JND(x,y,t)\right)^2 \delta(x,y,t))} \tag{6}$$

Authorized licensed use limited to: Peking University. Downloaded on November 23, 2009 at 05:14 from IEEE Xplore. Restrictions apply.