Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **5**, Number **2**, May **2020**

JND-based Multi-hypothesis Wyner-Ziv Video Coding

Jie Cheng[1], Lili Meng[1], Jia Zhang[1,2], Yanyan Tan[1], Yuwei Ren[1], Li Liu[1], Huaxiang Zhang[1]

[1]Department of Information Science and Engineering
Shandong Normal University, Jinan, China
[2]Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson,U.S.A
mengll_83@hotmail.com
Corresponding author: Lili Meng

Received February 2019; revised November 2019

Abstract. *Distributed Video Coding (DVC) plays a significant role in video coding system. Research communities think highly of it on how to improve the system performance. In this paper, we present a distributed video coding scheme based on just noticeable difference (JND). The JND is a measure of maximum image distortion that the human eye cannot detect due to the inherent characteristics of human vision. Therefore, JND model is applied to remove visual redundancy and reduce coding complexity. Meanwhile, the performance of distributed video coding system depends heavily on the quality of the side information (SI), and better performance can be expected when multiple SIs are utilized. Hence, the multi-hypothesis conditional probability density function (pdf) by Bayesian solution and weighted pdf are applied to the proposed scheme respectively. Experimental results illustrate that the proposed scheme is superior to other existing methods, and the Bayesian solution could outperform weighted pdf methods when better SIs are available.*
Keywords: Wyner-Ziv video coding; Multi-hypothesis; Just noticeable difference.

1. **Introduction.** In the most advanced video coding scheme like MPEG-4 [1] or the latest H.264 [2] standards, the encoder uses extremely complicated motion estimation algorithms to achieve efficient video compression, while the decoding process is low-complexity. However, applications such as wireless camera, burgeoning wireless video surveillance networks and mobile video telephone require low encoding complexity and high compression efficiency. DVC is a new paradigm which fits well these requirements since the complexity is shifted from the encoder to the decoder. The new paradigm is based on two major information theories: Slepian-Wolf (SW) lossless coding [3] and Wyner-Ziv (WZ) lossy coding [4]. DVC is an utterly advanced coding technique compared with the traditional video coding. In DVC system, the video sequence is first partitioned into even and odd frames. The even frames called W frames are intra-encoded and inter-decoded while the odd frames called K frames are encoded and decoded by conventional intra-frame coding methods. For each W frame, the decoder generates an estimation called side information (SI), from the previous decoded K frames. Then the channel decoder exploits the SI and the parity bits to reconstruct the W frame. As a result, the more accurate the SI is, the fewer parity bits are required from the encoder to decoder. In other words, if better quality SIs are obtained, the system performance will be enhanced. Thus, we can improve the compression performance of the system by improving the quality of SI.

One side information in reconstruction is developed in [5] under normal conditions. In [6], it is shown from the information theory perspective that the DVC coding efficiency can be enhanced by multiple SIs, because the conditional entropy of the source is the reduced. In the typical setting of the DVC, two SIs for each W frame can be readily obtained from the neighboring key frames, using forward and backward motion estimation respectively. In [7], a weighted average is proposed, and the weight is determined by the quality of the SI. In [8], it presents an effective method for getting better quality of two side information by using Bayesian conditional pdf. In a few words, if multiple SIs are obtained, the system performance can be improved.

JND refers to the characteristics of a variety of visual masking effects. Human visual system (HVS) is not sensitive to the information below JND threshold. Only the changes of the signal beyond the JND thresholds can be perceived by the human eye. In other words, the information below JND threshold is treated as visual redundancy which can be deserted and would not be encoded. The JND model is based on a psychological and physiological model. So far, several JND models have been proposed, which can be divided into two categories: JND model based on pixel domain [9] and JND model based on transform domain [10]. In the subsequent part, detail descriptions of the two models will be given. It is necessary to optimize coding algorithm using human visual characteristics on account of the human eye which is the final receiver. It can make the coding more consistent with human subjective experience by applying human visual characteristics to optimize the image coding algorithm.

The main innovation point of this paper is applying just noticeable difference (JND) model to multi-hypothesis distributed video coding system. JND model can remove visual redundancy which means the useless information would not be encoded. Thus, the coding efficiency of DVC could be enhanced. Also, the application of multiple SIs can improve the quality of reconstruction of video frames. Hence the system performance could be enhanced.

This paper is structured as follows. The proposed architecture is described in detail in Sec. 2. Sec. 3 describes the two JND model. The multi-hypothesis conditional pdf is shown in Sec. 4. Experiment and simulation results are shown in Sec. 5. Some summary is given in Sec. 6 finally.

2. **Overview of The Proposed System.**

2.1. **System Description.**

Fig. 1 illustrates the encoding and decoding process of the proposed JND-based Multi-hypothesis DVC. In the proposed scheme, the video sequence is divided into two categories firstly: W frames and K frames. And one W frame is between two K frames. A discrete cosine transform (DCT) with a 4×4 specification is done to W frames. The JND threshold of W frame in the DCT domain is then calculated. It is utilized to pretreat the DCT coefficient which is called filtering operation. If DCT coefficients value is lower than the JND threshold, it would be regarded as useless information and be abandoned. On the contrary, if the value is greater than the JND threshold, it would be retained. Also, the JND value in pixel domain can be obtained which is used to compute the PSPNR for the sake of making comparison between different schemes. The JND threshold in DCT domain and pixel domain will be discussed in detail in the later parts. After that procedure, the pretreated DCT coefficients are grouped together to form coefficients bands and then scalar quantized to form bit plane. The bit planes are sent to encoder to obtain parity bits and then encoded information of W frame will be sent to decoder. K frames are encoded and decoded by conventional video coding technology. After decoding the K frames, the two SIs are produced through forward prediction and

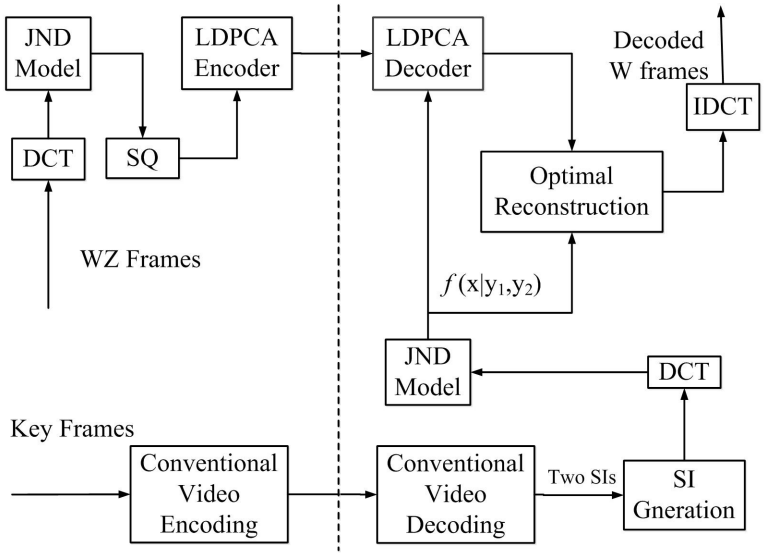

FIGURE 1. The architecture of proposed scheme.

backward prediction separately by using the decoded key frames. The two SIs are then transformed by DCT. JND threshold of side information in the DCT domain and pixel domain are also obtained, respectively. The same pretreatment is done to the DCT coefficients of the two SIs, and the conditional pdf used in ultimate reconstruction is obtained. The decoder combines side information with encoded information of W frame to make the reconstruction. After IDCT process, optimal reconstruction will be obtained ultimately. For the sake of testing performance of the proposed system, a series of experiments are carried out. In the latter portion, experimental results will be displayed and detailed analysis will also be given.

3. **JND Model.** JND refers to a measurement of maximum distortion of the human eye resulted from physiological and psychophysical phenomena in the human visual system. It is the minimum visibility threshold in essence when visual contents are altered. The application of JND promotes effective video compression [11], quality evaluation, watermarking. Human eye cannot distinguish the image changing because of some physiological limitation, and the visual redundancy can be reduced according to this characteristic. That means useless information that the human eye cannot detect can be decreased and only important information for the human eye can be treated. The JND model can be divided into two categories: JND model based on transform domain and JND model based on pixel domain. Here, the transform domain means the DCT domain. DCT coefficients of the original image are obtained firstly at the encoder, and then the DCT coefficients are pretreated by the DCT-based JND thresholds. The pixel domain JND thresholds, obtained by the DCT-based JND ones at the decoder, are used to handle the errors between the original image and the compressed image. The next is detailed descriptions of the two models.

3.1. **JND Model in DCT Domain.** The JND model in DCT domain is related to the spatial contrast sensitivity function (CSF), the luminance adaptation factor and the contrast masking weighting factor. And DCT-based JND model IS depicted in [10] as follows.

$$JND\_DCT(i,j) = J_{CSF}(i,j) \times F_{lum} \times F_{contrast}(i,j) \tag{1}$$

$J_{CFS}(i,j)$ represents the spatial contrast sensitivity function (CSF) and can be obtained from

$$J_{CSF}(i,j) = \frac{s}{\phi_i \phi_j} \times \frac{\frac{exp(c\omega_{ij})}{a+b\omega_{ij}}}{r + (1 - r\cos\psi_{ij})} \tag{2}$$

The value of a and b are 1.33 and 0.11. Experiential value of r is 0.6. s is measurement of spatial influence and usually its value is 0.25. ω_{ij} is spatial frequency of DCT coefficients. ϕ_i and ϕ_j are normalization coefficients of DCT. $r + (1 - rcos\psi_{ij})$ means tilting effect of human visual system. Meanwhile, ψ_{ij} means direction angle of corresponding DCT component.

F_{lum} is the luminance adaption factor. Due to adaptive luminance masking effect, human eye has the different sensitivity to different brightness areas of the image. The luminance adaptation factor can be achieved by equation (3).

$$F_{lum} = \begin{cases} 1 + \frac{60-\hat{I}}{150} & \hat{I} \leq 60 \\ 1 & 60 < \hat{I} < 170 \\ 1 + \frac{\hat{I}-170}{425} & \hat{I} \geq 170 \end{cases} \tag{3}$$

where $\hat{I}$ represents the average brightness and can obtained by

$$\hat{I} = \frac{DC}{N} + 128 \tag{4}$$

where N represent the size of image block and equals 4. DC is direct current coefficient in DCT domain. $F_{contrast}(i,j)$ is contrast masking weighting factor. Human eye have the different sensitivity to different areas of image so that different regions should be weighted different values. So, edge pixel density must be detected by edge detection on image for the sake of classifying the image block. ?$edgel$ is the proportion of edge pixels in the whole block and can be obtained from

$$\rho_{edgel} = \frac{\Sigma_{edgel}}{N^2} \tag{5}$$

N is block size and the value is 4. rho_{edgel} is the number of edge pixels in given block. The image blocks are divided by Canny edge detection into three categories: plane, edge, and texture. The classification principle is determined via (6).

$$type = \begin{cases} Plane & \rho_{edgel} \leq \alpha \\ Edge & \alpha < \rho_{edgel} \leq \beta \\ Texture & \rho_{edgel} > \beta \end{cases} \tag{6}$$

When $\alpha = 0.1$ and $\beta = 0.2$, the result is the best.

Ψ is weighted value, and the weighting principle of different regions is depicted by equation (7).

$$\Psi = \begin{cases} 1, & \text{for Plane and Edge block} \\ 2.25, & \text{for } (i^2 + j^2) \leq 16 \text{ in Texture block} \\ 1.25, & \text{for } (i^2 + j^2) > 16 \text{ in Texture block} \end{cases} \tag{7}$$

Considering the masking effect between the adjacent sub-bands, the contrast masking weighting factor is obtained:

$$F_{contrast}(i,j) = \begin{cases} \Psi, & \text{for } (i^2 + j^2) \leq 16 \text{ in Plane or Edge block} \\ \Psi \times \min(4, \max(1, (\frac{C(i,j)}{J_{CSF}(i,j) \times F_lum})^{0.36})), & \text{others} \end{cases} \tag{8}$$

After obtaining the threshold, the DCT coefficients of the image are pretreated. If the DCT coefficient is lower than the threshold, it would be abandoned. But if the DCT coefficient is larger than the threshold, it can be retained. Through such a process, redundant information is deleted. The advantage of doing this is to enhanced the coding efficiency.

3.2. **JND Model in Pixel domain.** Pixel-based JND estimation models have been developed and are often used in motion estimation, visual quality evaluation and video replenishment to avoid extra decomposition. Either the pixel domain JND model or the frequency domain JND model, they all come from the human visual mechanism. So they can be converted to each other.

The DCT coefficients of the original image are obtained by DCT transform. Then the JND threshold of the transform domain can be obtained. For the DC (direct current) and AC (alternating current) parts of the DCT coefficients will be generated after DCT conversion, the DC and AC parts also exist in the JND thresholds obtained by DCT coefficients. On the one hand, the DC part of the JND threshold is taken to estimate the local background brightness, and the local brightness threshold JND/N is obtained. On the other hand, the DCT coefficients of the image are processed by the JND threshold, and the new coefficients will be obtained. The processing method is shown below.

$$JND' = \begin{cases} sign(C) \cdot JND, & \text{if} \quad |C| \geq JND \\ 0, & \text{others} \end{cases} \tag{9}$$

where C represents the DCT coefficient. After this processing, the new coefficients are transformed by discrete cosine transform inverse transformation to get $Pixel\_JND'$. Finally, the final JND of the pixel domain is calculated as follows:

$$Pixel\_JND = \max\{Pixel\_JND', \frac{JND}{N}\} \tag{10}$$

After calculating the JND threshold of the DCT domain, the corresponding pixel domain JND threshold is subsequently calculated. The value is not processed for DCT coefficient, but is used to calculate PSPNR in the experimental part, which is used as a standard to evaluate the system performance.

4. **The Conditional pdf.** The quality of side information plays a definitive role in distributed video coding system. For different kind of side information, R-D performance of DVC system will be different. There already have been several methods to generate side information in DVC. Almost all the DVC codecs use the method of Motion Compensated Frame Interpolation (MCFI) to generate side information by interpolating current frame from the key frame. Then, it is proposed that iterative decoding and motion estimation can improve side information. Later, the idea of encoder aided motion estimation has been proposed to improve SI generation. In sake of enhancing the side information, Fan et al indicate approach of transform domain DVC in [12]. Then in [13], an improvement of transform-domain DVC was raised. Although it refines the quality of the SI after all the DCT bands are decoded, it raises decoding complexity. Up to now, an effective coding approach is the transform domain Wyner-Ziv video codec.

At the decoder of the distributed video coding system, there needs the statistical correlation model between SI and W frame. The relationship between SI and W frame can be described by $Y = X + Z$, where Y is side information and X stands for W frames. Z denotes the correlation noise. The conditional pdf of $f_{Y|X}(y|x)$ can be found equal to $f_{Y|X}(y|x) = f_Z(y - x)$. In many approaches, the Z is approximated by a Laplacian

distribution. In [5], the optimal reconstruction method of x using minimum mean-squared error (MMSE) has been proposed.

$$\hat{x}_{opt} = E[x|x \in [z_i, z_{i+1}], y_1, y_2] = \frac{\int_{z_i}^{z_{i+1}} x f_{X|y_1,y_2}(x)dx}{\int_{z_i}^{z_{i+1}} f_{X|y_1,y_2}(x)dx}. \tag{11}$$

So the conditional pdf $f_{X|y_1,y_2}(x)$ is needed at the decoder. Some existing methods have been proposed to approximate it. In [7], the following improved weighted average is proposed.

$$f_{X|y_1,y_2}(x) \approx w_1 f_{X|y_1}(x) + w_2 f_{X|y_2}(x), \tag{12}$$

where

$$w_i = \frac{\alpha_i^2}{\alpha_1^2 + \alpha_2^2}. \tag{13}$$

In [8], by invoking Bayes formula, the conditional pdf can be written as:

$$f(x|y_1, y_2) = \frac{f(x|y_1)f(x|y_2)f(y_1)f(y_2)}{f(x)f(y_1, y_2)}. \tag{14}$$

So the the optimal reconstruction can be expressed as

$$\hat{x}_{opt} = \frac{\int_{z_i}^{z_{i+1}} x \frac{f_{X|y_1}(x) f_{X|y_2}(x)}{f(x)} dx}{\int_{z_i}^{z_{i+1}} \frac{f_{X|y_1}(x) f_{X|y_2}(x)}{f(x)} dx}. \tag{15}$$

5. **Experiment and Results.**

5.1. **Evaluation Criterion.** In this section, the performance of our proposed scheme will be evaluated, and a series of experimental results will be shown. It should be noted that the peak signal-to-perceptual ration (PSPNR) in [14] is used as the criterion for evaluating the system performance. And the PSPNR is obtained from:

$$PSPNR = 10 \log_{10} \frac{255 \times 255}{\frac{1}{WH} \cdot \sum_{i=1}^{W} \sum_{j=1}^{H} (err(i,j)^2) \delta(i,j)} \tag{16}$$

and

$$\delta(i,j) = \begin{cases} 1, & |P(i,j) - \bar{P}(i,j)| \geq Pixel\_JND(i,j) \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where W and H denote the width and the height of the image. $Pixel_JND(i,j)$ represent the JND threshold of $pixel(i,j)$ of reconstructed frame. $P(i,j)$ and $P(i,j)$ represent the original value and the reconstructed value of the pixel located at (i,j), respectively.

5.2. **Experiment Results.** According to our proposed system, several experimental schemes are provided. The four test sequences whose R-D performance will be demonstrated are Foreman, Mother, Highway and Salesman. One chart is given for every test sequence for the sake of making comparisons between different schemes. Experimental results of Rate-Distortion of the proposed scheme are shown as followed.

Among these results, the system performance with two SIs scheme is superior to the system with one SI method which both of them are not applied JND method. That means if multiple side information is available, a better system performance can be acquired. Simultaneously, compared with one SI scheme, two SIs scheme with JND model also achieves a better effect.

Meanwhile, the charts present that our proposed method that using JND model in multi-hypothesis DVC system make a better R-D performance than the system without JND model. That means applying JND model in the multi-hypothesis DVC system is

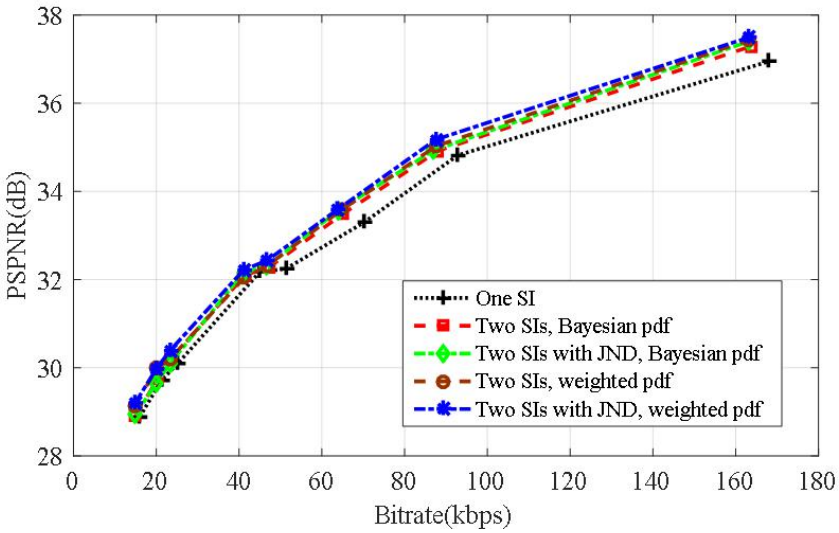

FIGURE 2. Rate-Distortion result for Foreman.

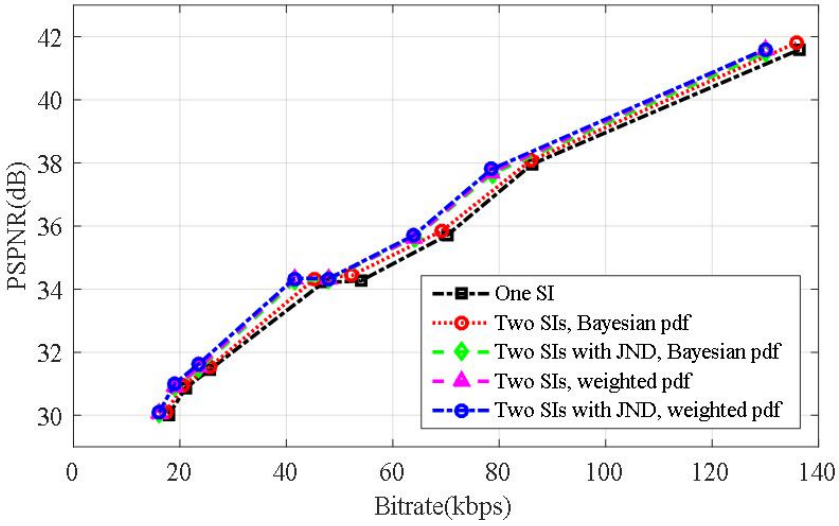

FIGURE 3. Rate-Distortion result for Mother.

feasible. However, when better SIs are available, how it will impact to different schemes with JND. For better simulating SIs of different qualities, the previous SIs are scaled as follows in [15]:

$$\begin{aligned} e_i &= y_i - x, \\ y_i' &= x + s * e_i, \end{aligned} \tag{18}$$

where the quality of the new SI is impacted by the scaling factor s. When $s = 1$, it means the SI is obtained by using forward or backward motion estimation.

Fig. 6 and Fig. 7 show the results of different methods with JND model for different scaling factors s. The test sequence are Foreman and Mother. But at the decoder, only the real SIs (when s =1) can be obtained. So that only average PSPNR of WZ frames at different bitrates are shown in Figure 8. and Figure 9.. It can be seen that when s ¡ 0.6, the system of Bayesian pdf method with JND model can outperform the system of Weighted pdf method with JND model. This shows the potential of the Bayesian approach.

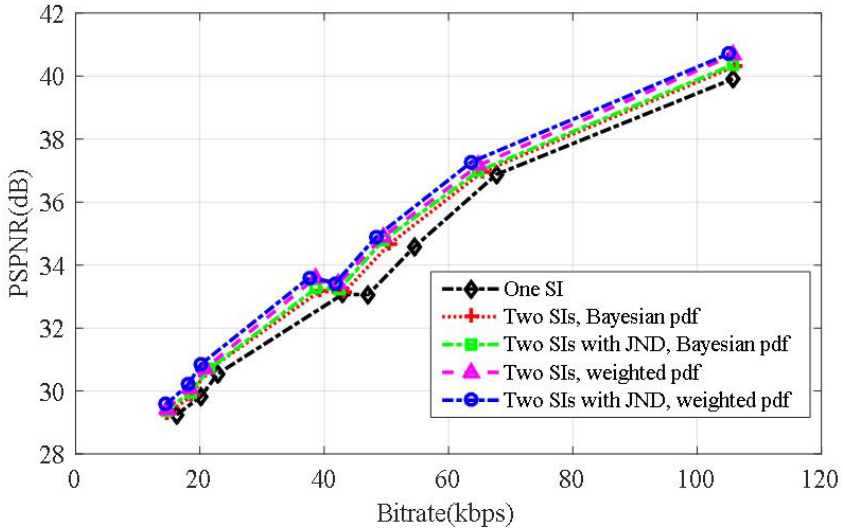

FIGURE 4. Rate-Distortion result for Salesman.

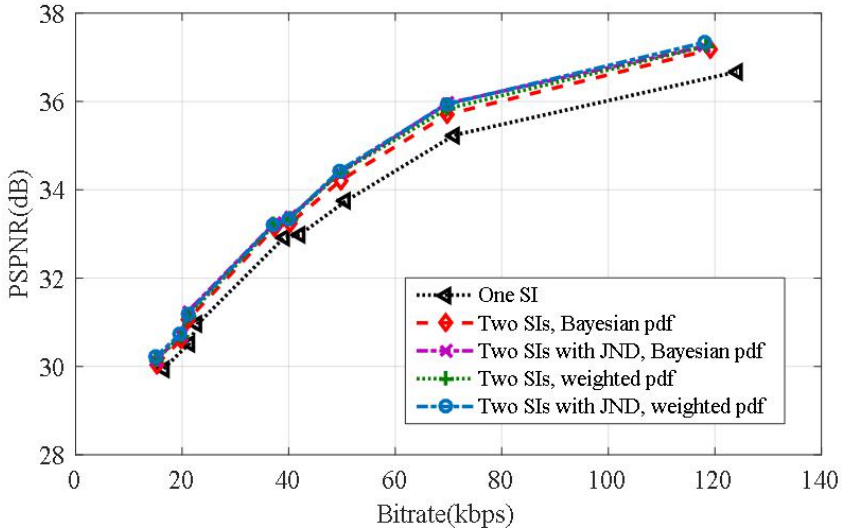

FIGURE 5. Rate-Distortion result for Highway.

6. **Conclusion.** In this paper, a new distributed video coding system called JND-based Multi-hypothesis distributed video coding is proposed. Human eye cannot distinguish the image changing because of some physiological limitation and can only apperceive the change of a specific threshold called just noticeable difference (JND). We only handle with the information greater than JND thresholds so that the coding efficiency of the system can be enhanced. In addition, our proposed system takes advantage of multiple side information. The multi-hypothesis pdf by Bayesian solution and weighted pdf are applied. Moreover, experimental results indicate the proposed scheme exceeds existing methods. Also, when better SIs are available, the Bayesian solution could outperform the Weighted pdf methods. What we will discuss next is how to obtain better quality of side information and reduce total encoding and decoding time.

Acknowledgment. The work is partially supported by the National Natural Science Foundation of China (No.61772322, U1736122, 61903238, 61702310).

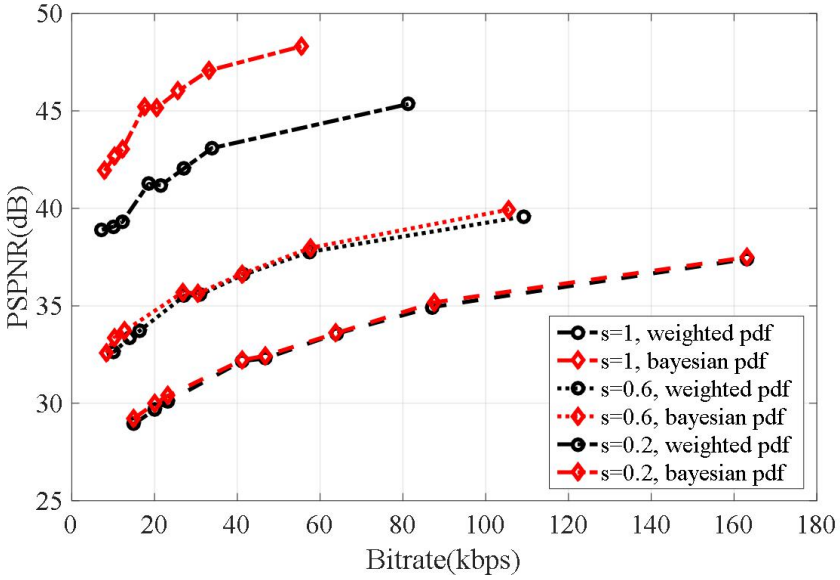

FIGURE 6. Rate-Distortion result using different-quality SIs in JND scheme for Foreman.

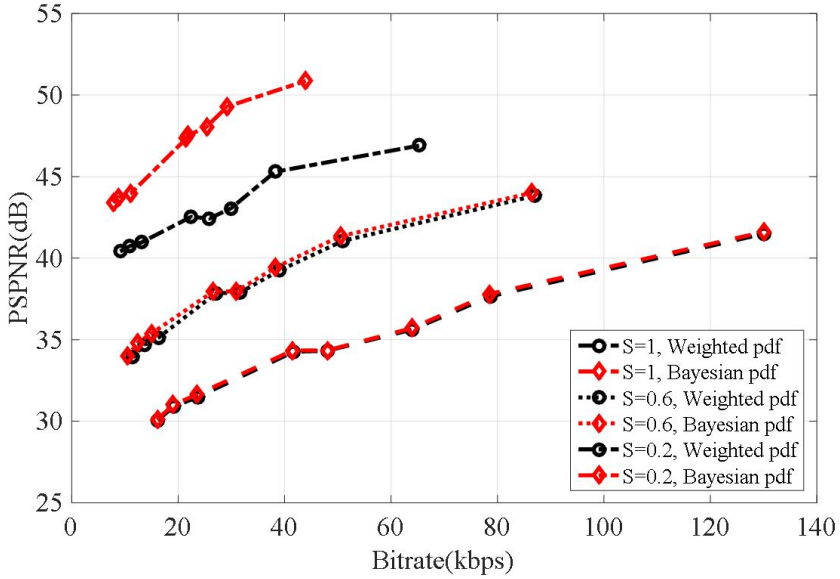

FIGURE 7. Rate-Distortion result using different-quality SIs in JND scheme for Mother.

REFERENCES

[1] SO/IEC International Standard 14496-2:2001/Amd 2. Information TechnologyCCoding of Audiovisual ObjectsCPart 2: Visual, Amendment 2: Streaming Video Profile.

[2] ISO/IEC International Standard 14496-10:2003. Information TechnologyCCoding of Audiovisual ObjectsCPart 10: Advanced Video Coding.

[3] J. D. Slepian and J. K. Wolf, Noiseless coding of correlated information sources, *IEEE Trans. on Information Theory*, vol. IT-22, pp. 471–480, Jul. 1973.

[4] A. D. Wyner and J. Ziv, The rate-distortion function for source coding with side information at the decoder, *IEEE Trans. on Information Theory*, vol. 22, pp. 1–10, Jan. 1976.

[5] D. Kubasov, J. Nayak and C. Guillemot, Optimal reconstruction in Wyner-Ziv video coding with multiple side information. *IEEE Multimedia Signal Processing Workshop,* October. 2007.

[6] K. Misra, S. Karande, and H. Radha, Multi-hypothesis distributed video coding using LDPC codes. *Proc, Allerton Conference on conmmunication, control and computing,* September. 2005.
[7] Y. Li, H. Liu, X. Liu, S. D. Ma, Zhao, and W. Gao, Multi-hypothesis based multi-view distributed video coding. *Picture Coding Symp.*, May, 2009.
[8] L.L. Meng and J.X. Zong, Bayesian Multi-Hypothesis Wyner-Ziv Video Coding. *Journal of Information Hiding and Multimedia Signal Processing* vol. 8, no. 2, March, 2017.
[9] X.K. Yang and W.S. Ling, Just noticeable distortion model and its applications in video coding. *Signal Processing Image Communication,* vol. 20, no. 7, pp. 662C680, August, 2005.
[10] Z. Wei and K.N. Ngan, Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain. *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, no. 3, pp. 337C346, March, 2009.
[11] X. Yang, W. Lin, Z. Lu, E. Ong and S. Yao, Rate control for videophone using local perceptual cues. *IEEE Transaction on Circuits and Systems for Video Technology,* vol. 15, no. 4, pp. 496C507, 2005.
[12] X. Fan, O. Au, and N. Cheung, Transform-domain adaptive correlation estimation (TRACE) for Video Technology. vol. 20, no. 11, pp. 1423-1436, 2010.
[13] R. Hansel and E. Muller, Improved adaptive temporal inter-/extrapolation schemes for distributed video coding. *Proceeding of the Picture Coding Symposium,* May 7-9, 2012, *Krakow, Poland. Piscataway, NJ, USA: IEEE,* pp. 213-216, 2012.
[14] C. Chou and C. Chou, A perceptually optimized 3-d subband codec for video communication over wireless. *IEEE Transaction on Circuits and Systems for Video Technology,* vol.6, no.2, 143-156, 1996.
[15] A. Aaron, S. Rane, and B. Girod, Transform-domain Wyner-Ziv codec for video. *VCIP,* pp. 520C528, January, 2004.