Multiple Description Video Coding Based on Human Visual System Characteristics

Huihui Bai, Weisi Lin, *Senior Member, IEEE*, Mengmeng Zhang, *Member, IEEE*, Anhong Wang, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— In this paper, a novel multiple description video coding scheme is proposed based on the characteristics of the human visual system (HVS). Due to the underlying spatial–temporal masking properties, human eyes cannot sense any changes below the just noticeable difference (JND) threshold. Therefore, at an encoder, only the visual information that cannot be predicted well within the JND tolerance needs to be encoded as redundant information, which leads to more effective redundancy allocation according to the HVS characteristics. Compared with the relevant existing schemes, the experimental results exhibit better performance of the proposed scheme at same bit rates, in terms of perceptual evaluation and subjective viewing.**

***Index Terms*— Just noticeable difference (JND), multiple description coding (MDC), video coding.**

I. Introduction

FOR compressed video, due to motion prediction and compensation, random bit errors and packet losses in transmission may cause substantial quality degradation. As a result, compressed video transmission over nonprioritized and unreliable networks is a challenging problem. Multiple description coding (MDC) is an attractive framework to tackle this problem. It can efficiently combat information loss without retransmission, thus satisfying the demand of real-time services and relieving the network congestion [1].

The original video signal can be split into multiple bit streams (descriptions) using a multiple description (MD) encoder. Then, these MDs can be transmitted over multiple channels. There is a very low probability when all channels fail at the same time. Therefore, at the MD decoder, only one description received can reconstruct the video with acceptable quality and the resultant distortion is called side distortion. Of course, more descriptions can produce the video with better quality. In a simple architecture of two channels, the distortion with two received descriptions is called central distortion [2].

During the past years, a number of methods of MDC have been presented, mainly including MD scalar quantizer [3], MD lattice vector quantizer [4], [5], MD based on pairwise correlating transforms [6], and MD based on FEC [7]. All the above methods have claimed to achieve good performance, but they are difficult to apply in practical applications because these specially designed MD encoders are not compatible with the widely used standard codec, such as H.26x and MPEG-x.

To overcome the limitation, some standard-compliant MD video coding schemes are designed to achieve promising results [8], [9]. Another significant class of MDC is based on pre- and postprocessing. In preprocessing, the original source is split into multiple subsources before encoding and then these subsources can be encoded separately by the standard codec to generate MDs. The typical versions are MDC based on spatial sampling [10] and temporal sampling [11], [12].

In the above mentioned schemes, effective redundancy allocation is the main task of MDC design. However, these schemes have not given full consideration to characteristics of the human visual system (HVS), which is the ultimate receiver of the decompressed video signal. In fact, human eyes cannot sense any changes below the just noticeable difference (JND) threshold around a pixel due to their underlying spatial–temporal masking properties, as explored in [13]. An appropriate (even imperfect) JND model can significantly help to improve the performance of video coding algorithms [14]. In [15], the JND model is applied in distributed video coding, and it is reported that the scheme can save the bit rates significantly without degrading the subjective quality of the reconstructed frames.

In this paper, the JND model is introduced to MDC for more effective redundancy allocation, as a way to base the proposed scheme with the convincing theoretical ground in perception science. In MDC, different descriptions can back-up or predict each other, when necessary (i.e., on occurrence of transmission loss), due to the built-in redundancy (i.e., a certain degree of predictability from other descriptions). Our idea in this paper, therefore, is: in view of effective redundancy allocation for the HVS characteristics, only the visual information that cannot be predicted well within the JND tolerance needs to be encoded as the redundant information. In this paper, we focus on the

Manuscript received May 28, 2013; revised October 28, 2013; accepted January 17, 2014. Date of publication April 4, 2014; date of current version August 1, 2014. This work was supported in part by the National Science Foundation of China under Grants 61272051, 61272262, 61073142, 61370111, and 61103113; in part by the Program for Changjiang Scholars and lnnovative Research Team in University under Grant IRT201206; in part by the 973 Program under Grant 2012CB316400; and in part by the Jiangsu Provincial National Science Foundation under Grant BK2011455. This paper was recommended by Associate Editor W. Zeng.

H. Bai and Y. Zhao are with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Institute Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: luckybhh@gmail.com; yzhao@bjtu.edu.cn).

W. Lin is with Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

M. Zhang is with the North China University of Technology, Beijing 100144, China (e-mail: zmm@ncut.edu.cn).

A. Wang is with the Taiyuan University of Science and Technology, Taiyuan 030024, China (e-mail: wah_ty@163.com).

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/TCSVT.2014.2315770

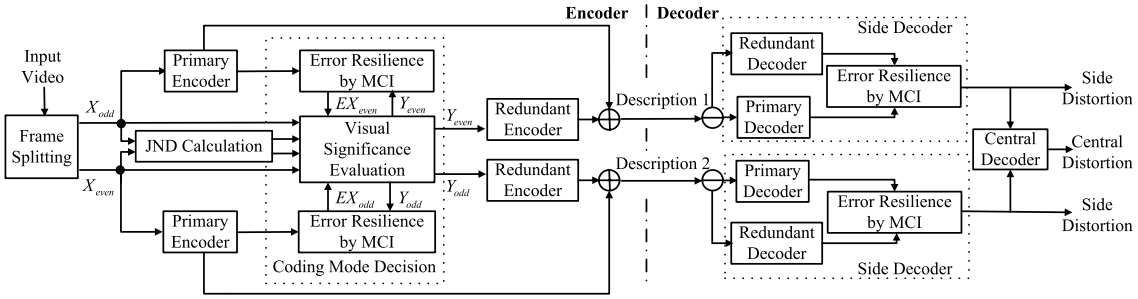


Fig. 1. Block diagram of the proposed scheme.

two-description MDC with odd and even frames, and the principles can be extended to MDs.

The rest of this paper is organized as follows. Section II presents the proposed scheme. In Section III, the performance of the proposed scheme is examined against the relevant existing MD coders. Finally, the conclusion is drawn in Section IV.

II. Proposed Scheme

A. Overview

Fig. 1 shows the block diagram of the proposed scheme. At the encoder, an original video sequence can be first split into odd and even frames, which are labeled as X_{odd} and X_{even}, respectively. For the two video subsequences X_{odd} and X_{even}, the standard video encoder can be used as primary encoder with full compliance. When one of the two compressed bit streams is lost, we can reconstruct it using the other bit stream, as a means of error resilience. For not introducing extra complexity to a standard coder, the same motion estimation and compensation method can be applied for error resilience, as the motion-compensated interpolation (MCI) [16] based on the piecewise uniform motion assumption to reconstruct the lost frames. Therefore, using the odd frames X_{odd}, we can obtain the estimated even frames EX_{even}, while using the even frames X_{even}, the estimated odd frames denoted by EX_{odd} can also be obtained.

According to [14], JND thresholds in spatial–temporal domain are then calculated to determine the visual sensitivity of the estimated odd and even frames, that is, EX_{odd} and EX_{even}. If the distortions from the estimated frames cannot be sensed by the HVS (i.e., below the JND thresholds), then no redundant information is needed for the description being considered. Otherwise, according to the JND thresholds, for the estimated frames EX_{odd} and EX_{even}, the needed redundancy Y_{odd} and Y_{even} can be encoded at the block level, respectively. Furthermore, adaptive coding mode decision for redundancy coding is performed toward achieving high compression efficiency. In the end, the bit streams from the primary encoders and redundant encoders form two descriptions, which can be transmitted over two channels.

At the decoder, if only one channel works, the side decoder is applied to obtain the reconstructed video with side distortion, while if two channels can work. Then the central decoder is employed to achieve the reconstructed video with central distortion. The important modules in Fig. 1 are to be further explained as follows.

B. Visual Significance Evaluation

In this module, according to the spatial–temporal JND thresholds [14], we determine whether the distortion of the estimated frames EX_{odd} and EX_{even} can be tolerated by the HVS. First, the spatial JND value of each pixel at (x, y) can be calculated as

$$\begin{aligned}&\text{JND}_s(x, y)\\ &\quad = T_l(x, y) + T_t(x, y) - C_{l,t} \cdot \min\{T_l(x, y), T_t(x, y)\} \quad (1)\end{aligned}$$

where the spatial JND thresholds are affected by two primary factors: 1) background luminance masking denoted by $T_l(x, y)$, which accounts for the fact that the HVS is more sensitive to luminance contrast rather than the absolute luminance value and 2) texture masking denoted by $T_t(x, y)$, which reflects that the HVS is more sensitive to the errors in smooth or edge regions than those in the textured regions. And $C_{l,t}$ $(0 < C_{l,t} < 1)$ reflects the compound effect for coexistence of luminance masking and texture masking; in this paper, $C_{l,t} = 0.3$ as in [14]. Furthermore, the spatial–temporal JND is then obtained by integrating temporal masking with $\text{JND}_s(x, y)$, which can be denoted as

$$\text{JND}(x, y, t) = f(ild(x, y, t)) \cdot \text{JND}_s(x, y) \quad (2)$$

where $ild(x, y, t)$ represents the average inter-frame luminance difference between tth and $t-1$th frame. And $f(.)$ is the function to account for the fact that usually a bigger inter-frame difference (caused by motion) leads to larger temporal masking [14]. The original video subsequences X_{odd} and X_{even} are used to calculate the JND thresholds $\text{JND}_i(x, y, t)$, using (1) and (2).

Next, the obtained JND thresholds can be employed to evaluate the visual quality of the estimated frames EX_{odd} and EX_{even}. Compared with the original video subsequence X_{odd} and X_{even}, the distortion of EX_{odd} and EX_{even} can be computed by

$$D_i(x, y, t) = |X_i(x, y, t) - EX_i(x, y, t)| \quad (3)$$

where i = odd or even. For each block, if the distortion $D_i(x, y, t)$ of $p\%$ pixels is larger than the corresponding spatial–temporal JND thresholds $\text{JND}_i(x, y, t)$, then adaptive coding should be used to include the perceptually significant

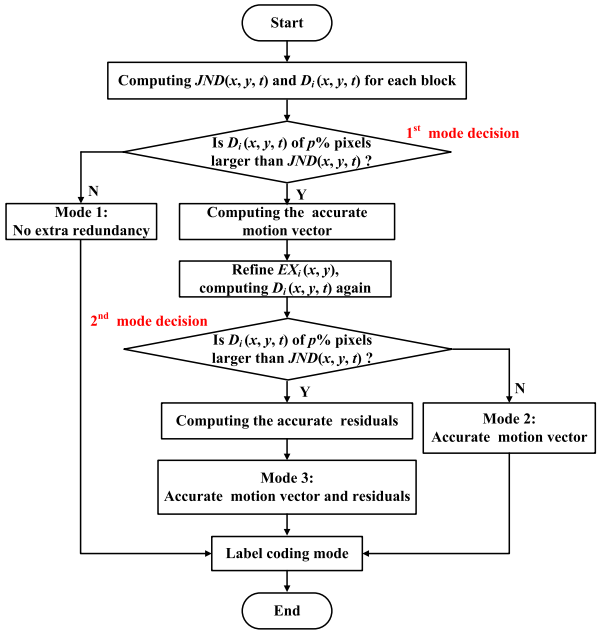


Fig. 2. Flowchart of coding mode decision.

redundancy as the descriptions. Otherwise, redundant information needs not to be coded since it is not perceptually significant.

C. Coding Mode Decision

According to the comparison between the pixels distortion $D_i(x, y, t)$ of each block in the estimated frames and the corresponding spatial–temporal JND thresholds $\mathrm{JND}_i(x, y, t)$, three modes of redundancy allocation will be designed as shown in Fig. 2.

For each image block in a description, if the distortion of $p\%$ pixels is smaller than the corresponding spatial–temporal JND thresholds, then the distortion from the error resilience can be considered perceptually acceptable by the HVS. In this case, no extra redundancy needs to be inserted, which can be regarded as Mode 1. On the other hand, if the distortion of $p\%$ pixels is larger than the corresponding spatial–temporal JND thresholds, then the HVS is considered sensitive to the errors from the estimated frames and the redundancy is needed. Here, we design other two coding modes for the redundancy allocation, as outlined below.

First, it is difficult for the MCI method to accurately estimate motion vectors in the decoder. Therefore, the motion vector for a block is regarded as the essential redundancy information, and is denoted by Mode 2. Such motion vectors can be searched in the original video subsequences in the module of visual significance evaluation at the encoder. Here, we use bidirectional motion estimation to find motion vectors. Then, the accurate motion vectors can be provided to the module of error resilience. Therefore, the blocks of estimated frames $\mathrm{EX}_{\mathrm{odd}}$ and $\mathrm{EX}_{\mathrm{even}}$ can be refined using these accurate motion vectors. The details can be found in Section II-D. If the distortion of these refined blocks is still larger than the spatial–temporal JND thresholds, then besides the motion vector like Mode 2, the corresponding residuals are also needed as additional redundancy. This mode can be considered as Mode 3.

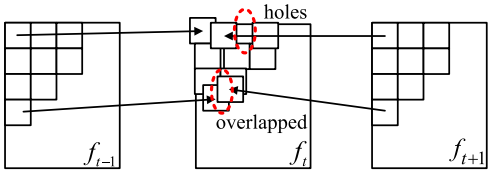


Fig. 3. Example of MCI.

D. Error Resilience by MCI

This module aims to achieve the corresponding estimated frames $\mathrm{EX}_{\mathrm{odd}}$ and $\mathrm{EX}_{\mathrm{even}}$ at the encoder, which can be then used for visual significance evaluation. We denote f_t as the estimated frame between frame f_{t-1} and frame f_{t+1} and $\mathrm{MV}(\vec{p})$ as the motion vector from f_{t-1} to f_{t+1} at the pixel location $\vec{p}$. Since both the two frames f_{t+1} and f_{t-1} are applied for forward and backward motion estimation and compensation, respectively, the overlapped pixels and holes may be produced in the reconstructed frame f_t, which is shown in Fig. 3.

To avoid the holes in the reconstructed frame f_t, we can first compute a preliminary reconstruction using

$$f_t(\vec{p}) = \frac{1}{2}(f_{t-1}(\vec{p}) + f_{t+1}(\vec{p})). \tag{4}$$

Furthermore, the forward and backward motion compensation is performed by frame f_{t+1} and f_{t-1}, respectively. To solve the overlapped problem, the mean values of overlapped pixels are adopted for the bidirectional motion compensation, that is, the preliminary reconstruction in (4) may be replaced according to

$$f_t(\vec{p}) = \frac{1}{2}\left(f_{t-1}\left(\vec{p}+\frac{1}{2}\mathrm{MV}(\vec{p})\right) + f_{t+1}\left(\vec{p}-\frac{1}{2}\mathrm{MV}(\vec{p})\right)\right). \tag{5}$$

In (5), due to the piecewise uniform motion assumption, the motion vector between frame f_t and frame f_{t+1} or f_{t-1} is considered as $(1/2)\mathrm{MV}(\vec{p})$.

It is noted that the MCI method mentioned previously is used for the initial estimated frame f_t in Fig. 2. After the first mode decision, the accurate motion vectors between frame f_t and frame f_{t-1} or f_{t+1} can be obtained as $\mathrm{MV}_{t-1,t}(\vec{p})$ or $\mathrm{MV}_{t,t+1}(\vec{p})$, which can be regarded as the redundancy information Y_{odd} and Y_{even} in Mode 2. Therefore, instead of (5), the blocks in the initial estimated frame f_t can be refined as

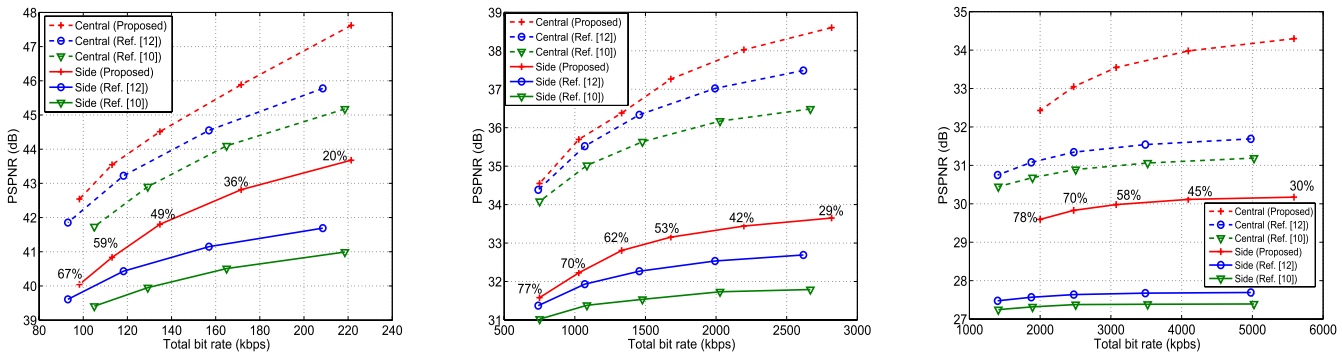
$$f_t(\vec{p}) = \frac{1}{2}(f_{t-1}(\vec{p} + \mathrm{MV}_{t-1,t}(\vec{p})) + f_{t+1}(\vec{p} - \mathrm{MV}_{t,t+1}(\vec{p}))). \tag{6}$$

After the second mode decision, the accurate residuals can also be achieved, which can be used to refine f_t in (6) further. The accurate motion vectors and residuals form the redundancy Y_{odd} and Y_{even} in Mode 3.

TABLE I

OVERALL SIDE AND CENTRAL QUALITY COMPARISON BETWEEN THE PROPOSED AND ANCHOR SCHEMES

Video sequence	$p\%$	Side PSPNR(dB)			Central PSPNR(dB)			Total bit rate(kbps)		Average ΔT
		Proposed	Ref. [12]	Gain	Proposed	Ref. [12]	Gain	Proposed	Ref. [12]	
Coastguard (QCIF)	56%	35.440	33.777	1.663	37.568	36.140	1.428	205.164	208.522	
Suzie (QCIF)	59%	40.835	40.431	0.404	43.544	43.222	0.322	113.033	118.243	5.89%
News (QCIF)	58%	43.388	40.722	2.667	46.947	45.796	1.150	124.736	138.169	
Mobile (CIF)	62%	32.806	32.267	0.539	36.381	36.335	0.046	1331.794	1456.841	
Paris (CIF)	63%	38.852	37.579	1.274	42.900	41.912	0.988	906.380	1029.146	6.15%
Soccer (4CIF)	70%	29.827	27.637	2.191	33.046	31.342	1.703	2474.172	2475.878	
Harbor (4CIF)	66%	37.762	36.732	1.030	41.241	40.597	0.644	4698.309	5487.499	7.02%
City (720P)	67%	36.603	34.897	1.706	38.669	37.336	1.333	2500.208	2603.815	8.37%
Average over 8 sequences		-	-	1.434	-	-	0.952	-	-	6.86%



(a) (b) (c)

Fig. 4. Performance comparison of side and central quality. (a) *Suzie*. (b) *Mobile*. (c) *Soccer*.

E. Decoder

Based on the above-mentioned encoder, we will discuss the decoder for the video reconstruction from two aspects: 1) side reconstruction and 2) central reconstruction.

If only one description is received, the side decoder is applied to obtain the side reconstruction. First, the bit stream of the received description is decoded by primary decoder and redundant decoder. Then, the module of error resilience same as the encoder is also employed estimate the loss information according to different coding decisions.

If two descriptions have been received, then the central decoder is used to reconstruct the video with central distortion. Here, the two reconstructed video from two descriptions can be exploited by averaging both representations to obtain a better reconstruction.

III. EXPERIMENTAL RESULTS

In the experiments, eight standard video sequences listed in Table I have been used to evaluate the efficiency of the proposed scheme, because they represent different visual content, motion, and resolution. The compared MDC schemes based on pre and postprocessing belong to the same class with the proposed scheme. In addition, the MDC scheme in [12] has also taken the visual quality into account, which presented the special design for scene change to obtain the better visual quality. To make a fair comparison, the same experimental setup and parameters have been chosen for the H.264 encoder and decoder with version JM10.2 of the software. Here, the same H.264 coding structure IPPP is adopted without B frames. And the same MCI is used in [12] for error resilience. It should be noted that the total bit rate is the sum of the two descriptions with the coding mode labels, and the subjective quality, i.e., the peak signal-to-perceptual ration (PSPNR) in [13] and [15] is applied to evaluate the reconstructed frames. PSPNR can be calculated as

$$\mathrm{PSPNR}(t) = 10\log_{10} \times \frac{255 \times 255}{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}(|I(x,y,t)-\hat{I}(x,y,t)|-\mathrm{JND}(x,y,t))^2\delta(x,y,t)} \tag{7}$$

and

$$\delta(x,y,t) = \begin{cases} 1, & \text{if } \left|I(x,y,t)-\hat{I}(x,y,t)\right| \geq JND(x,y,t) \\ 0, & \text{other} \end{cases} \tag{8}$$

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 results for all eight video sequences are listed in Table I. The average PSPNR gain is 1.434 dB for side reconstruction and 0.952 dB for central one, respectively. Furthermore, Fig. 4 shows a more extensive comparison at the different bit rates for video sequence *Suzie*, *Mobile*, and *Soccer*. From the results, the proposed scheme has an obvious advantage over the compared schemes in [10] and [12] in terms of the bit rate and visual distortion performance, especially at the higher bit rate. The main reason is that better redundancy allocation according

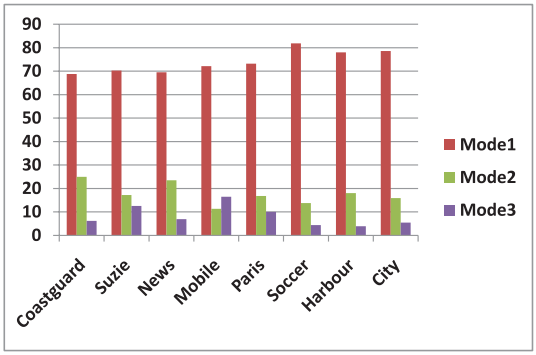


Fig. 5. Mode distribution for Table I.

(a) (b) (c) (d)

Fig. 6. Visual quality comparison for side reconstruction of *Soccer* 26th frame. (a) Original. (b) [12]: 2475.9 kb/s, PSPNR = 27.863 dB. (c) Proposed: 2474.2 kb/s, PSPNR = 30.172 dB, $p\%$ = 70%. (d) Proposed: 5588.3 kb/s, PSPNR = 31.098 dB, $p\%$ = 30%.

to the HVS is applied in the proposed scheme. In addition, with the increasing of bit rate, the redundancy allocation plays a more significant role. In Table I, the increasing computational load ΔT of the proposed scheme has also been presented ranging from 5.89% to 8.37% with different visual content; here, $\Delta T = (T_{\text{pro}} - T_{\text{ref}})/T_{\text{ref}}$, where T_{pro} and T_{ref} are the coding time for the proposed and [12] scheme, respectively.

It should be noted that in the proposed scheme the threshold $p\%$ is used to adjust the redundancy allocation adaptively. In mode decision, according to a given bit rate, the optimal $p\%$ can be determined to maximize the perceived quality. In Fig. 4, the optimal values of $p\%$ at different bit rate have been labeled and they are various adaptively according to visual content shown in Table I. From the results in Fig. 4, it can be seen that with the increasing of the bit rate, the values of $p\%$ have become smaller and smaller, which means more redundancy can be allocated adaptively. For the optimal $p\%$ in Table I, the distribution of each mode has been shown in Fig. 5.

Furthermore, from Fig. 6(b) and (c), at the similar bit rate, better visual quality has been observed in the proposed scheme than [12]. In addition, the perceived quality can be improved by adaptively down-adjusting the parameter $p\%$ for more redundancy allocation, as shown in Fig. 6(d) against (c), which means the distortion of smaller percentage pixels is larger than the JND thresholds.

IV. Conclusion

This paper proposes a novel idea for MD video coding based on the ground of the related HVS perception characteristics. The spatial–temporal JND model is applied for visual significance evaluation, and this leads to more efficient and user-centric redundancy allocation. To be more specific, only the perceptually significant visual information is included as redundant information. Experimental results have shown that the proposed scheme has achieved better bit rate and perceptual visual-distortion performance. Statistical redundancy has been substantially exploited for MDC, and we hope that the idea proposed in this paper serves as a starting point to explore perceptual redundancy more for the same.

References

[1] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, no. 1, pp. 57–69, Jan. 2005.
[2] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.
[3] C. Lin, Y. Zhao, and C. Zhu, "Two-stage diversity-based multiple description image coding," *IEEE Signal Process. Lett.*, vol. 15, no. 1, pp. 837–840, Aug. 2008.
[4] H. Bai, C. Zhu, and Y. Zhao, "Optimized multiple description lattice vector quantization for wavelet image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 7, pp. 912–917, Jul. 2007.
[5] M. Liu and C. Zhu, "M-description lattice vector quantization: Index assignment and analysis," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2258–2274, Jun. 2009.
[6] A. R. Reibman, H. Jafarkhani, and Y. Wang, "Multiple-description video coding using motion-compensated temporal prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 3, pp. 193–204, Mar. 2002.
[7] S. Chang, P. C. Cosman, and L. B. Milstein, "Performance analysis of channel symmetric FEC-based multiple description coding for OFDM networks," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1061–1076, Apr. 2011.
[8] C. Lin, T. Tillo, Y. Zhao, and B. Jeon, "Multiple description coding for H.264/AVC with redundancy allocation at macro block level," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 589–600, May 2011.
[9] Y. Xu and C. Zhu, "End-to-end rate-distortion optimized description generation for H.264 multiple description video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1523–1536, Sep. 2013.
[10] A. Wang, Y. Zhao, and H. Bai, "Robust multiple description distributed video coding using optimized zero-padding," *Sci. China Ser. F, Inf. Sci.*, vol. 52, no. 2, pp. 206–214, Feb. 2009.
[11] C. Zhu and M. Liu, "Multiple description video coding based on hierarchical B pictures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 4, pp. 511–521, Apr. 2009.
[12] M. Zhang and H. Bai, "Adaptive multiple description video coding and transmission for scene change," *EURASIP J. Wireless Commun. Netw.*, vol. 265, pp. 1–3, Aug. 2012.
[13] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Just noticeable distortion model and its application in video coding," *Signal Process., Image Commun.*, vol. 14, pp. 662–680, Aug. 2005.
[14] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, Jun. 2005.
[15] Y. Li, D. Zhao, S. Ma, and W. Gao, "Distributed video coding based on the human visual system," *IEEE Signal Process. Lett.*, vol. 16, no. 11, pp. 985–988, Nov. 2009.
[16] Y. Wang, J. Ostermann, and Y. Zhang, *Video Processing and Communication*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.