Multidim Syst Sign Process (2017) 28:1249–1266
DOI 10.1007/s11045-016-0395-2

Perceptual feature guided rate distortion optimization for high efficiency video coding

Aisheng Yang[1] · Huanqiang Zeng[1] · Jing Chen[1] · Jianqing Zhu[2] · Canhui Cai[2]

Received: 19 December 2015 / Revised: 7 February 2016 / Accepted: 8 March 2016 /
Published online: 16 March 2016

Abstract With the advances in understanding perceptual properties of the human visual system, perceptual video coding, which aims to incorporate human perceptual mechanisms into video coding for maximizing the perceptual coding efficiency, becomes an essential research topic. Since the newest video coding standard—high efficiency video coding (HEVC) does not fully consider the perceptual characteristic of the input video, a perceptual feature guided rate distortion optimization (RDO) method is presented to improve its perceptual coding performance in this paper. In the proposed method, for each coding tree unit, the spatial perceptual feature (i.e., gradient magnitude ratio) and the temporal perceptual feature (i.e., gradient magnitude similarity deviation ratio) are extracted by considering the spatial and temporal perceptual correlations. These perceptual features are then utilized to guide the RDO process by perceptually adjusting the corresponding Lagrangian multiplier. By incorporating the proposed method into the HEVC, extensive simulation results have demonstrated that the proposed approach can significantly improve the perceptual coding performance and obtain better visual quality of the reconstructed video, compared with the original RDO in HEVC.

Keywords Human visual system · High efficiency video coding · Perceptual feature · Rate distortion optimization

1 Introduction

In recent years, high definition (HD) and ultra HD (UHD) videos are being increasingly popular due to their high fidelity, which poses great challenge on the video storage and transmission. In this context, the joint collaborative team on video coding (JCT-VC) standard organization formed by video coding experts group (VCEG) and moving picture experts

✉ Huanqiang Zeng
zeng0043@hqu.edu.cn

1 School of Information Science and Engineering, Huaqiao University, Xiamen 361021, China

2 School of Engineering, Huaqiao University, Quanzhou 362021, China

group (MPEG) developed the newest video coding standard, namely, high efficiency video coding (HEVC) Sullivan et al. (2012), Ugur et al. (2010). Compared with the previous video coding standards, HEVC still follows the classical hybrid block-based coding framework and introduces a lot of the latest video technological achievements Wang et al. (2014). With much higher coding efficiency, HEVC is expected to be widely applied in different areas, such as online education, video broadcasting, entertainment, and so on.

As we known, the objective of developing the video coding technology is to provide the highest *perceptual* visual quality under a given bit rate budget. However, similar to the previous video coding standards, HEVC still exploits the classical Lagrangian rate distortion optimization (RDO) technique to maximize the coding efficiency measured by the objective criterion (i.e., Sum of Square Error, SSE), which can not accurately reflect the perceptual visual quality Girod (1993). In other words, HEVC does not fully consider the perceptual characteristics of the input video during the encoding process and thus has room to be improved in the perceptual measurement. Since the video quality is ultimately judged by human eye, it is very desirable to develop some perceptual guided optimization strategies for HEVC in order to improve its perceptual coding efficiency.

First of all, with the development in understanding perceptual properties of the human visual system (HVS), visual quality assessment (VQA) has attracted more and more attentions from both academical and industrial communities Wang et al. (2004), Zhang et al. (2011), Ma et al. (2011), Xue et al. (2014). For example, the well-known VQA metric—structural similarity index (SSIM), which can accurately describe the structure inherited in visual content, has been proposed by Wang et al. (2004). Zhang et al. (2011) proposed a practical full-reference (FR) metric by jointly considering the texture masking effect and contrast sensitivity function. Ma et al. (2011) proposed a reduced-reference (RR) image quality assessment based on the statistical model of the discrete cosine transform (DCT) coefficient distribution. For the perceptual video coding, the straightforward solution is to incorporate the existing VQA metrics that can more accurately describe the human perception into the video codec to improve its perceptual coding efficiency. For example, some perceptual video coding methods utilized the well-known SSIM in the DCT domain Wang et al. (2013), RDO Huang et al. (2010), Wang et al. (2012), Yeo et al. (2013) and rate control Ou et al. (2011), Zhao et al. (2013) to reduce the perceptual redundancies. Among them, Wang et al. (2013) proposed a normalization factor based on DCT domain-SSIM index to transform the DCT residual into the perceptually uniform space together with a new distortion model for improve the perceptual coding efficiency. Huang et al. (2010) introduced the SSIM to replace the SSE as the quality metric and developed a SSIM-based RDO using the coding information of the key frame. Yeo et al. (2013) further derived the relationship between the SSIM-based RDO and the original RDO used in H.264/AVC. Ou et al. (2011) presented a rate control method by exploiting the SSIM-based RDO based on the framework presented in Huang et al. (2010). Zhao et al. (2013) proposed an improved Largest coding unit (LCU)-level rate control algorithm for HEVC based on the SSIM. In this method, the SSIM is used to decide the weight of LCU-level bit allocation in the R-λ model so that the rate is allocated based on the perceptual characteristics of each LCU.

However, it should be pointed out that some VQA metrics are not applicable to perceptual video coding due to high computational complexity and compatibility issues. Different from the above-mentioned methods that optimize the coding performance in terms of the existing VQA metric—SSIM, some perceptual video coding methods exploited various important perceptual properties of HVS to measure the perceptual characteristics in video, such as visual attention, visual sensitivity, contrast sensitivity, structural information, to name a few Chen et al. (2010), Lee and Ebrahimi (2012). Xu et al. (2014) proposed a region-of-interest (ROI)

based HEVC coding approach for conversional videos with a novel hierarchical perception model of face. Moreover, a weight-based unified rate-quantization scheme is proposed to adaptively adjust the value of quantization parameter (QP) rather than the conventional pixel-based unified rate-quantization scheme. Meddeb et al. (2014) developed a new rate control scheme for HEVC standard with the aim of improving the perceptual quality of ROI. Li et al. (2015) suggested a new weight-based R-λ method for the rate control in HEVC specifically for conversational videos by incorporating the perceptual properties of the conversational videos. Wang et al. (2014) proposed a Lagrangian multiplier based perceptual optimization scheme to improve the perceptual quality for HEVC. Zeng et al. (2013) proposed an adaptively adjusting Lagrangian multiplier in the RDO process based on the perceptual sensitivity of the input CTU. Jung and Chen (2015) obtained an adaptive Lagrangian multiplier based on the free-energy principle which represents the disorderly concealment effect in human eyes. Zeng et al. (2015) proposed a perceptual rate control method for HEVC to obtain the perceptual coding gain by adaptively allocating the bits for the region with different perceptual measurements. Xu et al. (2013) proposed a new visual quality metric, in which MSE is weighted spatially and temporally to simulate the HVS response to visual signal. Moreover, the visual quality metric related to quantization parameter is capable of guiding perceptual video coding. Ma et al. (2011) proposed a novel adaptive block-size transform based just-noticeable different (JND) model to improve the perceptual coding performance by considering both the spatial and temporal perceptual features. Kim et al. (2015) introduced a HEVC-compliant perceptual video coding scheme based on the JND models in both transform and pixel domains for variable block-size transform kernels. The transform-domain JND model is designed by adopting an existing pixel-domain JND model for the transform skip model and considering the spatial JND characteristics.

In this paper, a perceptual feature guided rate distortion optimization (RDO) approach is proposed for HEVC. The improvement of the perceptual coding efficiency is due to that the proposed method adaptively adjusts the Lagrangian multiplier in the RDO process according to the perceptual characteristic of video content, which is evaluated by two perceptual features extracted for each CTU. As a result, the CTU with higher texture complexity or lower perceptual distortion will be allocated with less bits, since it can tolerate more distortion. On the contrary, more bits will be allocated to the CTU that is more sensitive in human perception. Simulation results show that the proposed method can significantly improve the perceptual coding performance, compared with the original RDO in HEVC.

The remaining parts of this paper is organized as follows. The RDO conducted in the HEVC is briefly introduced in Sect. 2. The proposed perceptual feature guided RDO method is presented in Sect. 3. The simulation results are provided in Sect. 4. Finally, the concluding remarks are given in Sect. 5.

2 Rate distortion optimization in HEVC

In the HEVC, rate distortion optimization (RDO) plays an important role in the mode decision process, which is to find a good trade-off between the reconstructed video quality and the required bits. To adapt to various video content, there are various prediction modes in HEVC that can be roughly classified into Intra, SKIP, and Inter Modes, as shown in Fig. 1. For example, the mode with larger size (e.g., $M \times M$) consumes less bits for head information and is efficient to code the picture block with homogeneous textures. In contrast, the mode with smaller size (e.g., $M/2 \times M/2$) provides more accurate prediction and thus yields less residual, at the expense of higher head information. The mode decision process is to exhaustively

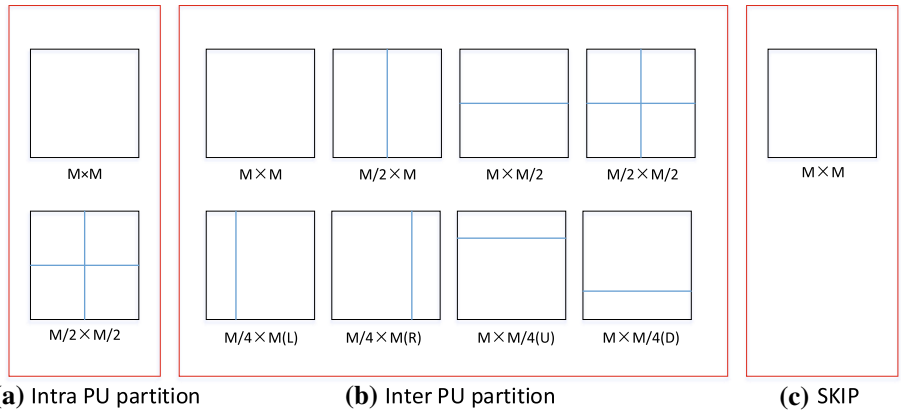


Fig. 1 Prediction modes in HEVC

compute the RD cost of all the prediction modes and find the one with the minimum RD cost as the optimal mode. In fact, this process is an optimization problem that minimizes the overall reconstructed video distortion D at a given rate R.

The Lagrangian RDO used in HEVC is defined as:

$$J_{RD} = D + R \cdot \lambda_{HEVC} \tag{1}$$

where J_{RD} is the Lagrangian cost function, D means the distortion between the original block and its reconstructed block, R represents the total number bits for coding the headers, quantized coefficients, etc. measured in terms of bits per pixel. The λ_{HEVC} is the Lagrangian multiplier, which is a weighting factor between distortion and bits as defined:

$$\lambda_{HEVC} = \alpha \cdot 2^{\left(\frac{QP-12}{3.0}\right)} \tag{2}$$

where QP is the quantization parameter, and the α is a constant that is empirically-determined and defined in HEVC.

It can be easily observed from formulation (1) that the RDO in HEVC ignores the perceptual characteristics of the video content. On one hand, the distortion is measured in terms of SSE that is not highly related to the HVS. On the other hand, the λ_{HEVC} plays a very important role in the optimization of the coding performance. For example, a lager λ_{HEVC} will result in a higher distortion and a lower bit rate, and vice versa. Unfortunately, one can see from Eq. (2) that the λ_{HEVC} is only a function of QP, which does not consider the video contents and their perceptual characteristics. The above analysis indicates that the HEVC is not so efficient in the sense of perceptual video coding. To improve the perceptual coding efficiency, perceptual RDO is an efficient solution.

3 Proposed perceptual feature guided RDO for HEVC

It is known that the perceptual video coding is to remove the perceptual spatial and temporal redundancies inherited in the video. Hence, the key of perceptual video coding becomes how to extract the perceptual feature and how to use the related perceptual feature to guide the video coding. In this work, to improve the perceptual coding efficiency, a perceptual feature

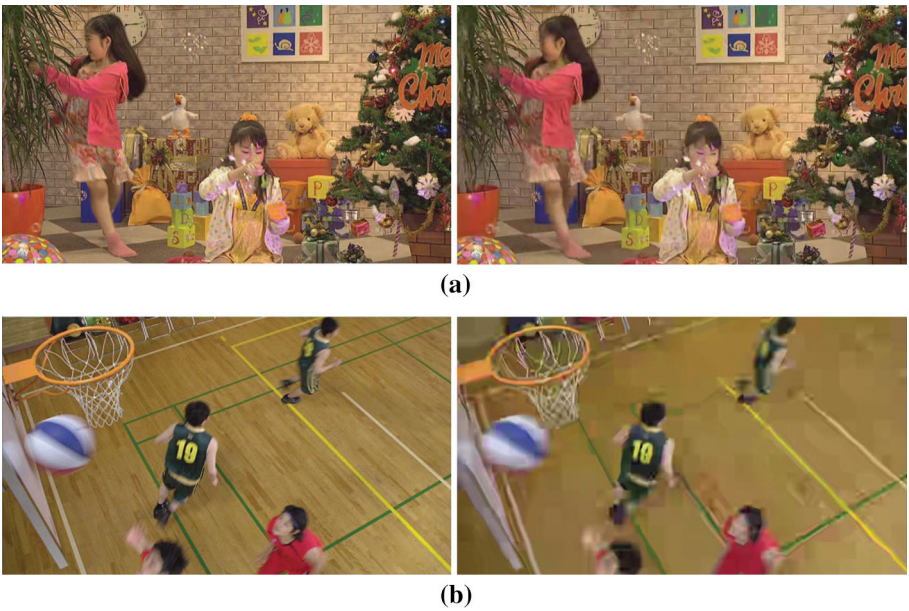


Fig. 2 An example of the reconstructed video frame with different texture complexity and similar PSNR under low delay setting in HEVC: **a** "PartyScene" (1920 × 1080, 11th frame, $PSNR = 26.31$ dB); **b** "BasketballDrill" (1920 × 1080, 2nd frame, $PSNR = 26.59$ dB), where the original frame and the reconstructed frame are listed from left to right

guided RDO is proposed for HEVC by perpetually adjusting Lagrangian multiplier based on the perceptual features of the input video content. More specifically, the perceptual features are extracted and utilized to guide the adjustment of the Lagrangian multiplier so that the bit rates can be adaptively allocated based on the perceptual features of the video content. Consequently, the perceptual coding efficiency is greatly improved.

3.1 Perceptual features

First, from the spatial viewpoint, the viewer is more sensitive to the smooth texture region; on the contrary, the complex texture region is detail-irrelevant and thus quite a large amount of errors can be hidden in such kind of region Lee and Ebrahimi (2012). In the other words, the same distortion will produce higher perceptual quality reduction in the smooth texture region than the complex texture region. Figure 2 shows an example of the reconstructed video frame with complex texture and smooth texture under the similar PSNR in HEVC. One can see that sequence "PartyScene" in Fig. 2a contains multiple objects and sophisticate and messy background, which is more complex than sequence "BasketballDrill" in Fig. 2b. By introducing the similar distortion (i.e., with similar PSNR) produced by HEVC into these two different kinds of video frames, it can be observed that the perceptual quality of the video frame "PartyScene" is acceptable while that of "BasketballDrill" is significantly reduced, for example, the face of the player becomes blur and the block artifacts are very obvious in the basketball court. This study clearly indicates that the complex texture region is able to hide more distortion than the smooth region. Further study shows that the smoothness of the image region can be reflected by its edge information Xue et al. (2014), Tang et al. (2006). It

means that the human perception of the image region can be indicted by its edge description. Based on this intuition, the gradient is used to extract the spatial perceptual feature due to its simplicity and efficiency. Since the basic processing unit in HEVC is the coding tree unit (CTU), the gradient computation is performed on each CTU to measure its smoothness. More specifically, the CTU with higher gradient magnitude tends to have complex texture while the one with lower gradient magnitude tends to be a smooth CTU.

In this work, for each CTU, the gradient extraction is to convolve it with the Prewitt filter H, along the horizontal (i.e., H_x) and vertical (i.e., H_y) directions as below:

$$H_x = \frac{1}{3} \times \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}, \quad H_y = \frac{1}{3} \times \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \tag{3}$$

For the i-th pixel (i.e., P_i) in the CTU, its gradient magnitude, namely, GM_{P_i}, can be computed as follows:

$$GM_{P_i} = \sqrt{(CTU \otimes H_x)^2(i) + (CTU \otimes H_y)^2(i)} \tag{4}$$

where the symbol $\otimes$ means the convolution operation. Consequently, the total gradient magnitudes of the current CTU are computed by adding the gradient magnitudes of all the pixels as follows:

$$GM_{CTU} = \sum_{i=1}^{N} GM_{P_i} \tag{5}$$

where N is the number of pixel in the CTU. Similarly, the mean gradient magnitude of each frame is computed by:

$$MGM_{CTU} = \frac{\sum_{i=1}^{K} GM_{CTU}(i)}{K} \tag{6}$$

where K is the number of CTU in a frame. Intuitively, the mean gradient magnitude represents the average degree of the texture complexity. If the CTU has higher gradient magnitude than this mean gradient magnitude value, it means that this CTU tends to have more complex texture in this frame. Therefore, a gradient magnitude ratio (GMR) is presented to describe the perceptual feature of the current CTU, which can be computed as:

$$GMR(i) = \frac{GM_{CTU}(i) + c_1}{MGM_{CTU} + c_1} \tag{7}$$

where c_1 is an empirically-determined small constant to avoid the instable special case that the current CTU is very smooth and thus the mean gradient magnitude is zero. One can see that the smaller the GMR value is, the smoother the current CTU is. On the contrary, the more complex CTU will correspond to higher GMR value.

Second, from the temporal viewpoint, there exists strong temporal corrections between the current CTU and its temporal adjacent CTUs. Moreover, the temporal adjacent CTUs have been coded and some coding information can be reused. Hence, it would be reasonable to use the perceptual characteristic of the temporal adjacent CTUs to predict that of the current CTU. Motivated by this, instead of the commonly-used objective quality metrics (e.g., SSE, MAD), the recent proposed well-known subjective quality assessment—gradient magnitude similarity deviation (GMSD) Xue et al. (2014) is exploited to measure the perceptual quality of the temporal-adjacent CTUs in the first stage. Then, the temporal perceptual feature of the current CTU, $GMSD_{CTU_0}$, can be predicted by using the temporal correlation as follows. Figure 3 shows the current CTU, CTU_0 and its temporal adjacent CTUs.

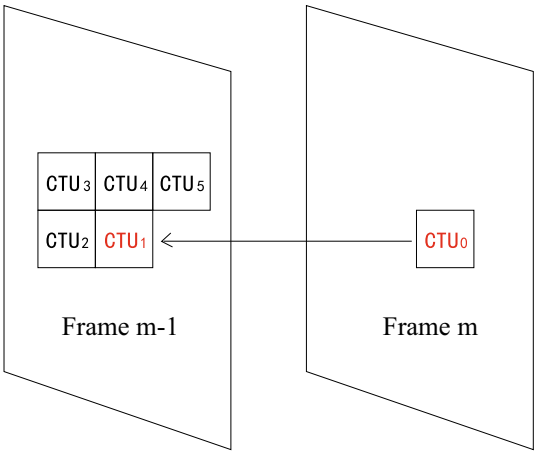


Fig. 3 The current CTU, CTU_0 and its temporal adjacent CTUs

Table 1 The weights w_i for the temporal adjacent CTUs

$CTU's$ index i	1	2, 3, 4, 5
w_i	0.5	0.125

$$GMSD_{CTU_0} = \sum_{i=1}^{5} GMSD_{CTU_i} \cdot w_i \tag{8}$$

where w_i are the weights of the corresponding CTU_i in Fig. 3, for i=1, 2, 3, 4 and 5. It is known that the closer the neighboring CTU is, the more similar the neighboring CTU is, the larger weight the neighboring CTU is. Based on this intuition, the weights w_i are empirically determined from extensive experiments and are shown in Table 1. The $GMSD_{CTU_i}$ means the GMSD of corresponding CTU_i, where the GMSD has been demonstrated its effectiveness on perceptual quality assessment Xue et al. (2014) and can be computed as:

$$\begin{aligned} GMSD &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GMS(i) - GMSM)^2} \\ GMSM &= \frac{1}{N}\sum_{i=1}^{N} GMS(i) \\ GMS(i) &= \frac{2GM_r(i)GM_d(i) + c}{GM_r(i)^2 + GM_d(i)^2 + c} \end{aligned} \tag{9}$$

where N means the number of pixel in current CTU, $GMSM$ means the mean of the gradient magnitude similarity of all the pixels in the current CTU, $GMS(i)$ means the gradient magnitude similarity of current CTU at pixel i, the $GM_r(i)$ and $GM_d(i)$ represent the gradient magnitudes of the reference and distorted CTUs at pixel i, which can be computed according to Eq. (4).

Note that, for the current CTU, the higher the GMSD score is, the lager the perceptual distortion is, which means this CTU could tolerate less distortion. Similarly, the GMSD ratio, $GMSDR$, is presented to describe the perceptual ratio of the current CTU and can be computed as:

$$GMSDR_{CTU_0} = \frac{GMSD_{CTU_0}}{MGMSD} \tag{10}$$

where

$$MGMSD = \frac{\sum_{i=1}^{K} GMSD_{CTU_i}}{K} \tag{11}$$

where K is the number of CTUs in a frame, $MGMSD$ is the mean GMSD of all the CTUs in the temporal adjacent frame. It can be easily observed that the higher the $GMSDR$ value is, the less perceptual distortion can be added. On the contrary, the CTU with the lower $GMSDR$ tends to tolerate more distortion.

3.2 Perceptual feature guided RDO

Based on the above analysis, the perceptual features GMR and $GMSDR$ can effectively reflect the perceptual characteristic of the current CTU. The consequent problem is how to use these perceptual features to guide the encoding process for improving the perceptual coding performance. In this work, a perceptual feature guided RDO is presented by adaptively adjusting the Lagrangian multiplier for each CTU based on the extracted perceptual features as follows.

The perceptual features GMR and $GMSDR$ represent the perceptual characteristic of the current CTU from the spatial and temporal viewpoints, respectively. On one hand, the larger GMR means the current CTU tends to have more complex texture. On the other hand, the smaller $GMSDR$ means the current CTU tends to have good perceptual quality already. Hence, the CTU with larger GMR and lower $GMSDR$ should be allocated with less bits. Intuitively, they contribute equally to the perceptual evaluation of the current CTU. Hence, a simple strategy for the evaluation of the perceptual characteristic (PC) of the current CTU is to combine them together by multiplying GMR and the inverse of $GMSDR$, as follows.

$$PC = \frac{GMR}{GMSDR} \tag{12}$$

One can see that the PC value will be increased with the values of GMR and $\frac{1}{GMSDR}$, which makes full use of the spatial and temporal perceptual correlations of the current CTU and its adjacent CTUs. The larger PC value indicates that the current CTU has more complex texture or less perceptual distortion, which can be hidden with more distortion and thus be allocated with less bits. Hence, it is reasonable to allocate the bits based on the PC value of the current CTU. Moreover, from the RDO process referred to Eq. (1), the larger λ_{HEVC} will make the current CTU choosing the mode that produces higher distortion and lower bit rate. Therefore, the perceptual characteristic PC can be incorporated into the RDO process by guiding the adjustment of the Lagrangian multiplier to adaptively allocate the bits for each CTU so that the perceptual RD performance can be improved. To be more specific, the CTU with larger PC value can be assigned with the larger Lagrangian multiplier, as it can tolerate larger amount of distortions. On the contrary, CTU with smaller PC value can be assigned with the smaller Lagrangian multiplier so that more bits can be allocated with these more sensitive CTUs. In the other words, the relationship between the Lagrangian multiplier and PC value is monotone increasing. Hence, the proposed perceptual feature guided RDO can be summarized as follows:

$$J_{RD} = D + R \cdot \lambda_{PC} \tag{13}$$

where λ_{PC} is the perceptual feature guided Lagrangian multiplier and can be defined as below:

$$\lambda_{PC} = (a \cdot PC + b) \times \lambda_{HEVC} \tag{14}$$

where a and b are empirically determined as 1.2 and 0.01 from extensive experiments, respectively.

In summary, the proposed perceptual feature guided RDO algorithm for HEVC can be described as below:

1) For the current frame, compute the gradient magnitude of each CTU according to (5) and then obtain its mean gradient magnitude MGM according to (6).
2) For the current CTU, compute the spatial perceptual feature—gradient magnitude ratio GMR based on (7).
3) Estimate the gradient magnitude similarity deviation $GMSD$ of the current CTU according to (8) and then obtain the temporal perceptual feature—gradient magnitude similarity deviation ratio $GMSDR$ based on (10).
4) Compute the perceptual characteristic (PC) according to (12) and then derive the perceptual feature guided λ_{PC} by (14).
5) Proceed to the mode decision process of the current CTU using the proposed perceptual feature guided RDO.
6) Repeat Steps (2) to (5) until all the CTUs in the current frame are encoded.
7) Repeat Steps (1) to (6) until all the frames are encoded.

4 Experiment result and discussion

4.1 Test conditions

To validate the performance of the proposed perceptual feature guided RDO scheme, it has been integrated into the HEVC reference software (i.e., HM10.0 HM (2013)). All test video sequences are in YCbCr 4:2:0 format with various resolutions and video contents. The performance is tested under the standard Low Delay with IBBB structure setting and Random Access setting Bossen (2012). Moreover, the RDO is enabled and the quantization parameter (QP) value is set as 22, 27, 32, and 37.

To evaluate the perceptual RD performance, some commonly-used well-known perceptual quality metrics---SSIM Wang et al. (2004), Gradient Magnitude Similarity Mean (GMSM), Gradient Magnitude Similarity Deviation (GMSD) Xue et al. (2014) are used to measure the perceptual quality of the reconstructed video instead of the traditional objective metric—PSNR, as they have been demonstrated their superiority on the perceptual quality assessment. Note that the higher SSIM and GMSM values and the lower GMSD value indicate the better perceptual quality. Similar to PSNR, the perceptual quality of the video is obtained by simply averaging the corresponding perceptual metric value computed on each frame. Furthermore, the proposed method is compared with the original RDO in the HEVC. The average difference between their perceptual rate-distortion (RD) curves is measured according to the method in Bjontegaard (2001), the performance index ΔBR is used to measure the total bit rate changes (in percentage) under the same perceptual distortion measured in terms of different perceptual quality metrics (i.e., SSIM, GMSM, GMSD), and the performance indexs $\Delta SSIM$, $\Delta GMSM$, $\Delta GMSD$ are used to measure the perceptual quality changes under the same bit rates.

Table 2 The perceptual RD performance comparison in terms of SSIM

Sequence	Resolutions	Low delay (%)		Random access (%)	
		$\Delta BR(\%)$	$\Delta SSIM$	$\Delta BR(\%)$	$\Delta SSIM$
BasketballPass	416 × 240	−7.00	0.0046	−7.05	0.0045
RaceHorses	416 × 240	−1.94	0.0015	−1.93	0.0015
BQSquare	416 × 240	−10.06	0.0046	−6.77	0.0036
BlowingBubbles	416 × 240	−1.40	0.0009	−2.16	0.0016
BQMall	832 × 480	−4.32	0.0011	−3.12	0.0009
PartyScene	832 × 480	−4.58	0.0012	−2.78	0.0012
BasketballDrill	832 × 480	−8.72	0.0038	−6.75	0.0029
RaceHorses	832 × 480	0.25	−0.0005	0.12	0.0003
FourPeople	1280 × 720	−5.08	0.0005	−1.53	0.0002
KristenAndSara	1280 × 720	0.27	0	−1.62	0.0002
BasketballDrive	1920 × 1080	−7.69	0.0012	−6.57	0.0011
Kimonol	1920 × 1080	−4.17	0	−1.35	0.0001
ParkScene	1920 × 1080	−3.34	0.0009	−3.52	0.0009
BQTerrace	1920 × 1080	−17.11	0.0009	−12.68	0.0007
Cactus	1920 × 1080	−8.19	0.0006	−23.07	0.0008
Traffic	2560 × 1600	−6.77	0.0009	−10.88	0.0008
PeopleOnStreet	2560 × 1600	−15.34	0.0012	−8.94	0.0008
Average		−6.18	0.0014	−5.92	0.0013

4.2 Perceptual RD performance comparison

Tables 2 and 3 and 4 individually show the performance of the proposed perceptual feature guided RDO method for the HEVC in terms of SSIM, GMSM and GMSD, compared with the original RDO in HEVC. Moreover, the perceptual RD curves of some test sequences "BQTerrace" (1920 × 1080) and "BQSquare" (416 × 240) under the low delay and random access settings are shown in Figs. 4 and 5 as examples, respectively. One can see from these tables and figures that the proposed perceptual feature guided RDO method is able to achieve a much better perpetual RD performance under both low delay and random access settings, compared with the original RDO in HEVC. To be more specific, by maintaining the same perceptual quality measured in terms of SSIM, GMSM and GMSD, 6.18 %, 12.72 %, 19.80 % bit rate reduction for low delay setting and 5.92 %, 9.92 %, 16.24 % bit rate reduction for random access setting can be achieved by the proposed method. Meanwhile, with the same bit rate, the proposed method can obtain 0.0014 SSIM, 0.0074 GMSM improvement and 0.0096 GMSD decrement for low delay setting and 0.0013 SSIM, 0.0058 GMSM increment and 0.0078 GMSD decrement for random access setting. Moreover, to have a clearer demonstration, Figs. 6 and 7 show two examples of the reconstructed video frame by the original RDO in HEVC and the proposed method under low delay setting and random access setting, respectively. It can be easily observed that compared with the original RDO in HEVC, the proposed method can effectively reduce the bits while keeping the similar perceptual quality.

Table 3 The perceptual RD performance comparison in terms of GMSM

Sequence	Resolutions	Low delay (%)		Random access (%)	
		$\Delta BR(\%)$	$\Delta GMSM$	$\Delta BR(\%)$	$\Delta GMSM$
BasketballPass	416×240	−14.49	0.0124	−13.18	0.0108
RaceHorses	416×240	−4.05	0.0025	−3.52	0.0021
BQSquare	416×240	−17.22	0.0147	−9.48	0.0115
BlowingBubbles	416×240	−3.15	0.0013	−3.14	0.0012
BQMall	832×480	−9.52	0.0047	−7.07	0.0034
PartyScene	832×480	−5.74	0.0022	−5.76	0.0024
BasketballDrill	832×480	−10.62	0.0074	−8.50	0.0060
RaceHorses	832×480	−7.25	0.0043	−7.49	0.0037
FourPeople	1280×720	−18.19	0.0072	−11.01	0.0055
KristenAndSara	1280×720	−13.62	0.0081	−13.53	0.0073
BasketballDrive	1920×1080	−10.52	0.0055	−11.49	0.0059
Kimonol	1920×1080	−6.11	0.0029	−6.16	0.0030
ParkScene	1920×1080	−18.49	0.0122	−12.59	0.0084
BQTerrace	1920×1080	−34.85	0.0142	−24.10	0.0105
Cactus	1920×1080	−12.45	0.0054	−10.39	0.0052
Traffic	2560×1600	−11.32	0.0106	−11.36	0.0063
PeopleOnStreet	2560×1600	−18.72	0.0099	−10.77	0.0095
Average		−12.72	0.0074	−9.92	0.0058

Table 4 The perceptual RD performance comparison in terms of GMSD

Sequence	Resolutions	Low delay (%)		Random access (%)	
		$\Delta BR\,(\%)$	$\Delta GMSD$	$\Delta BR\,(\%)$	$\Delta GMSD$
BasketballPass	416×240	−15.70	−0.0124	−13.83	−0.0176
RaceHorses	416×240	−5.43	−0.0032	−4.42	−0.0026
BQSquare	416×240	−19.28	−0.0175	−10.80	−0.0139
BlowingBubbles	416×240	−4.89	−0.0022	−3.17	−0.0015
BQMall	832×480	−12.82	−0.0058	−7.89	−0.0037
PartyScene	832×480	−7.41	−0.0035	−8.14	−0.0042
BasketballDrill	832×480	−12.49	−0.0078	−9.64	−0.0063
RaceHorses	832×480	−9.82	−0.0041	−19.71	−0.0110
FourPeople	1280×720	−26.60	−0.0091	−19.00	−0.0075
KristenAndSara	1280×720	−51.83	−0.0103	−49.04	−0.0085
BasketballDrive	1920×1080	−14.26	−0.0058	−14.40	−0.0059
Kimonol	1920×1080	−11.18	−0.0042	−10.27	−0.0039
ParkScene	1920×1080	−34.85	−0.0019	−20.91	−0.0122
BQTerrace	1920×1080	−50.72	−0.0256	−35.62	−0.0173
Cactus	1920×1080	−20.59	−0.0075	−20.27	−0.0071
Traffic	2560×1600	−15.01	−0.0123	−12.75	−0.0071
PeopleOnStreet	2560×1600	−23.69	−0.0123	−13.22	−0.0106
Average		−19.80	−0.0096	−16.24	−0.0078

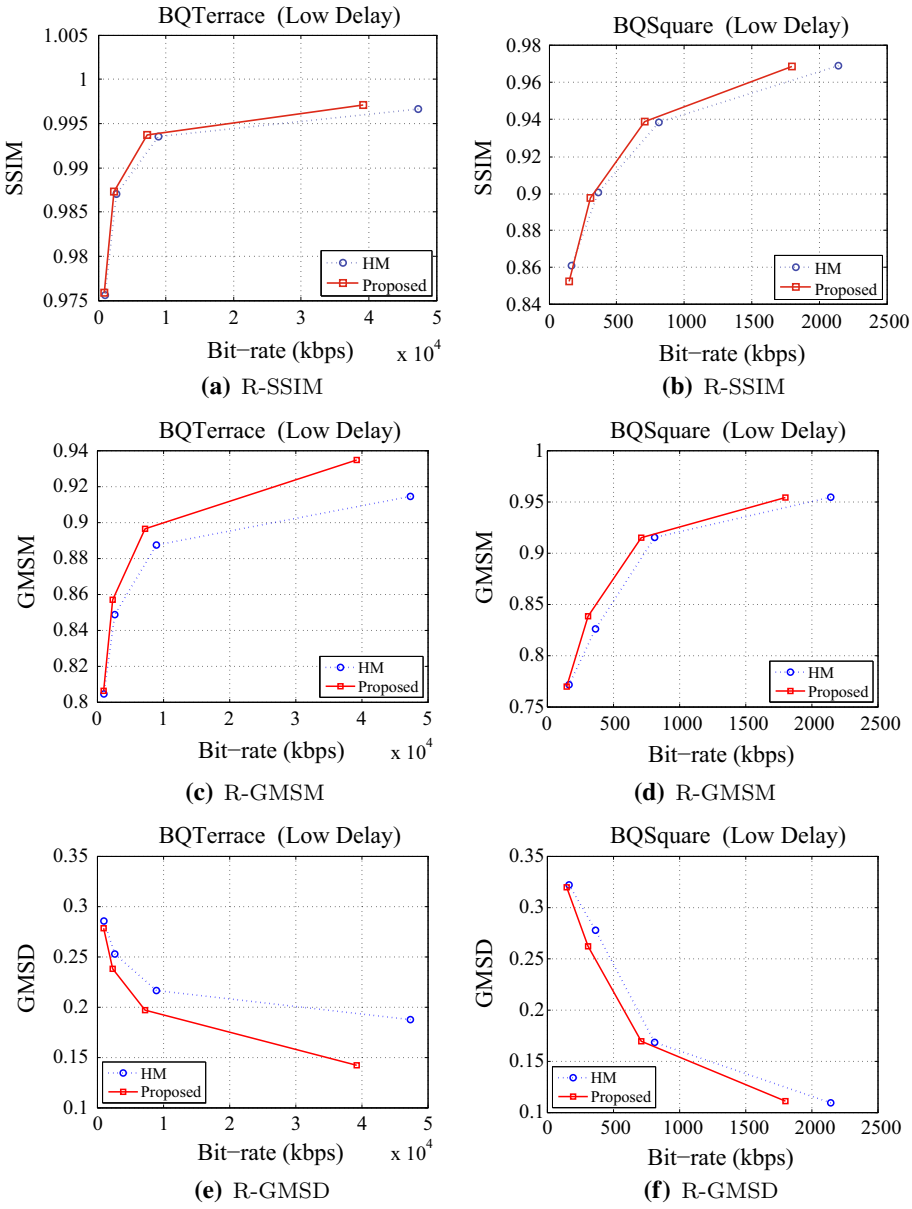


Fig. 4 The perceptual RD curves of test sequences "BQTerrace" and "BQSquare" under low delay setting

4.3 Discussion

To further explain the logical behind the proposed method, taking the sequence "BasketballPass" under the random access setting as an example, Fig. 8 shows the λ value used by the original RDO in HEVC and the proposed method, and the corresponding reconstructed video frames. One can see that the λ value in the RDO of HEVC is the same for all the CTUs,

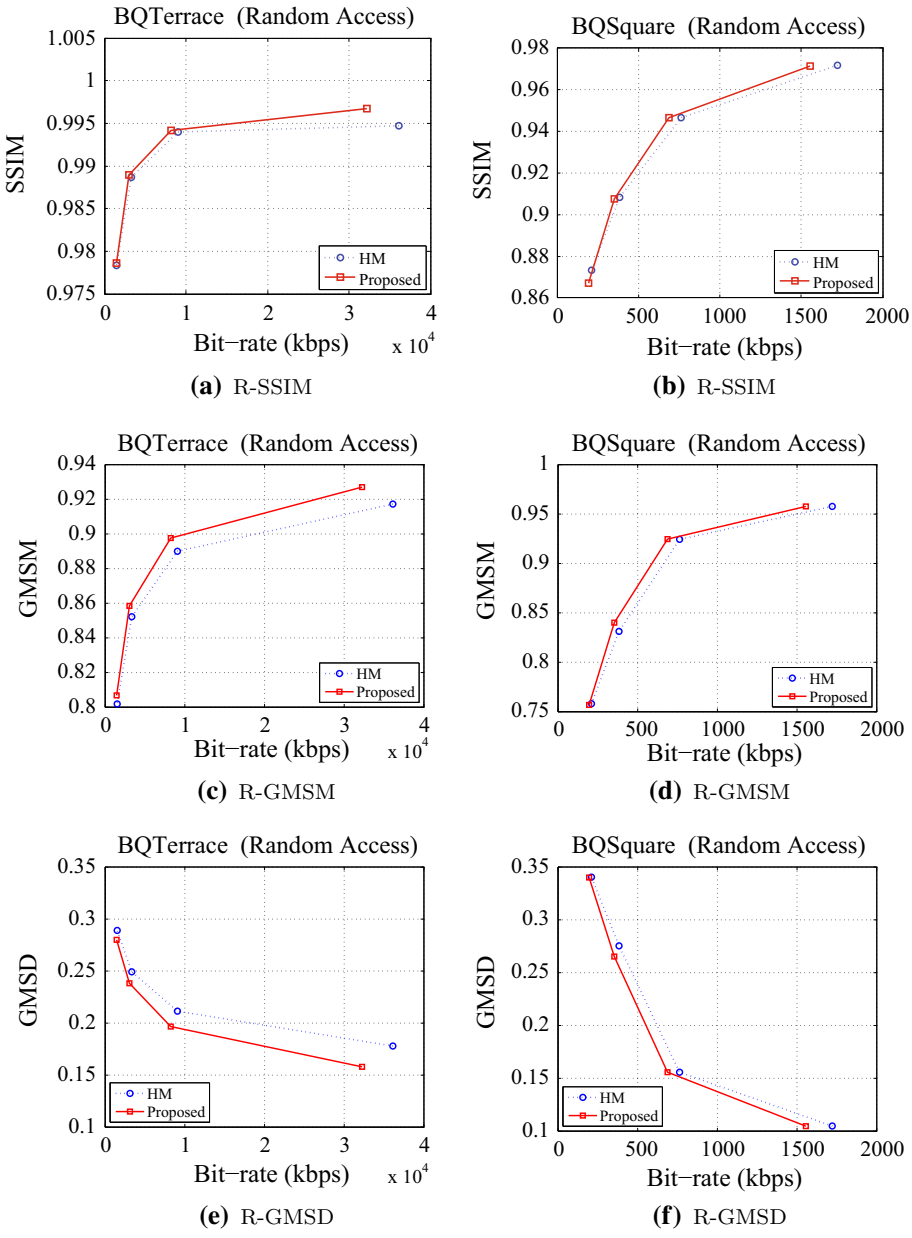


Fig. 5 The perceptual RD curves of test sequences "BQTerrace" and "BQSquare" under random access setting

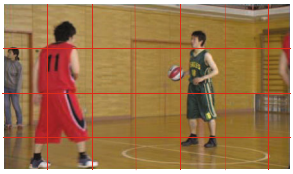
which is only related to the QP as shown in (2). It means that the RDO in HEVC ignores the perceptual characteristics of the video content. Hence, to improve the perceptual coding efficiency, the proposed perceptual feature guided RDO takes the perceptual characteristic of each CTU into account during the encoding process. To be more specific, the perceptual

(a) (b)

Fig. 6 The reconstructed video frames—"Cactus" (23rd frame, QP = 22) under low delay setting **a** original HEVC (SSIM: 0.9956, GMSM: 0.9005, GMSD: 0.1694, bits: 260,504 bits); **b** proposed method (SSIM: 0.9952, GMSM: 0.9056, GMSD: 0.1762, bits: 221,536 bits)

(a) (b)

Fig. 7 The reconstructed video frames—"BasketballPass" (9th frame, QP = 22) under random access setting: **a** original HEVC (SSIM: 0.9747, GMSM: 0.9610, GMSD: 0.0949, bits: 4568 bits); **b** proposed method (SSIM: 0.9742, GMSM: 0.9609, GMSD: 0.096, bits: 3480 bits)



(a)

5.61	5.61	5.61	5.61	5.61	5.61	5.61
5.61	5.61	5.61	5.61	5.61	5.61	5.61
5.61	5.61	5.61	5.61	5.61	5.61	5.61
5.61	5.61	5.61	5.61	5.61	5.61	5.61

(b)

(c)

1.09	0.94	0.64	0.91	1.29	1.22	1.19
1.46	1.00	0.62	0.73	2.10	0.63	1.27
1.05	1.29	0.57	0.43	1.78	0.71	1.34
1.14	0.66	0.48	0.69	1.26	0.87	0.52

(d)

7.36	6.34	4.35	6.13	8.71	8.25	8.05
9.87	6.73	4.16	4.91	14.15	4.29	8.60
7.12	8.72	3.83	2.91	12.03	4.82	9.06
7.71	4.44	3.27	4.71	8.53	5.89	3.51

(e)

(f)

Fig. 8 Sequence "BasketballPass"(416×240, 2nd frame) under random access setting: **a** the original video frame; **b** λ value used in original HEVC; **c** the reconstructed video frame by the original HEVC (GMSD: 0.09248, bits: 41,736 bits); **d** the perceptual characteristic *PC* value computed by the proposed method; **e** λ value used in the proposed method; **f** the reconstructed video frame by the proposed method (GMSD: 0.09244, bits: 38,960 bits)

characteristic of each CTU will be firstly evaluated according to Eq. (12). Then, the CTU with larger PC value, which is less perceptual sensitive, will be assigned with larger λ to allocate less bits, and vice versa. This can be further verified by Fig. 8 that the proposed method is able to assign different λ value to different CTUs according to their PC values. Consequently, the perceptual RD performance by the proposed method is effectively improved, compared with the original RDO in HEVC.

5 Conclusion

In this paper, a perceptual feature guided RDO is proposed for HEVC to improve its perceptual coding performance. In our approach, the perceptual features for each CTU are firstly extracted and then integrated into the RDO process by perceptually adjusting the Lagrangian multiplier. The perceptual coding gain by the proposed method is achieved by the adaptive bit allocation for each CTU based on its perceptual characteristic. Experimental results show that the proposed method is able to obtain significant improvement on perceptual RD performance and visual quality of the reconstructed video, compared with the original RDO in HEVC.

Acknowledgements This work was support in part by the National Natural Science Foundation of China under the Grants 61401167 and 61372107, in part by the Natural Science Foundation of Fujian Province under the Grant 2016J01308, in part by the Opening Project of State Key Laboratory of Digital Publishing Technology under the Grant FZDP2015-B-001, in part by the Zhejiang Open Foundation of the Most Important Subjects, in part by the High-Level Talent Project Foundation of Huaqiao University under the Grants 14BS201 and 14BS204, and in part by the Graduate Student Scientific Research Innovation Ability Cultivation Plan Projects of Huaqiao University under the Grant 1400201031.

References

Bjontegaard, G. (2001). Calculation of average PSNR differences between RD-curves (VCEG-M33). In *VCEG meeting (ITU-T SG16 Q. 6)*.

Bossen, F. (2012). Document JCTVC-J1100: Common test conditions and software reference configurations. In *JCT-VC Meeting*, Stockholm, Sweden, Tech. Rep.

Girod, B. (1993). What's wrong with mean-squared error? In Andrew B. Watson *Digital images and human vision* (pp. 207–220). MIT Press, Cambridge.

Girod, B. (1993). What's wrong with mean-squared error? In A. B. Watson (Ed.), *Digital images and human vision* (pp. 207–220). Cambridge: MIT Press.

Bjontegaard, G. (2001). Calculation of average PSNR differences between RD-curves (VCEG-M33). In *VCEG meeting (ITU-T SG16 Q. 6)*.

Huang, Y. H., Ou, T. S., Su, P. Y., & Chen, H. H. (2010). Perceptual rate-distortion optimization using structural similarity index as quality metric. *IEEE Transactions on Circuits and Systems for Video Technology*, *20*(11), 1614–1624.

Jung, C., & Chen, Y. (2015). Perceptual rate distortion optimisation for video coding using free-energy principle. *Electronics Letters*, *51*(21), 1656–1658.

Kim, J., Bae, S. H., & Kim, M. (2015). An HEVC-compliant perceptual video coding scheme based on JND models for variable block-sized transform kernels. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(11), 1786–1800.

Lee, J. S., & Ebrahimi, T. (2012). Perceptual video compression: A survey. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 684–697.

Li, S., Xu, M., Deng, X., & Wang, Z. (2015). Weight-based R-λ rate control for perceptual HEVC coding on conversational videos. *Signal Processing: Image Communication, 38*, 127–140.

Ma, L., Li, S., Zhang, F., & Ngan, K. N. (2011). Reduced-reference image quality assessment using reorganized DCT-based image representation. *IEEE Transactions on Multimedia*, *13*(4), 824–829.

Ma, L., Ngan, K. N., Zhang, F., & Li, S. (2011). Adaptive block-size transform based just-noticeable difference model for images/videos. *Signal Processing: Image Communication*, *26*(3), 162–174.
Meddeb, M., Cagnazzo, M., & Pesquet-Popescu, B. (2014). Region-of-interest-based rate control scheme for high-efficiency video coding. *APSIPA Transactions on Signal and Information Processing*, *3*, e16.
Ou, T. S., Huang, Y. H., & Chen, H. H. (2011). SSIM-based perceptual rate control for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, *21*(5), 682–691.
Sullivan, G. J., Ohm, J. R., Han, W. J., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(12), 1649–1668.
Tang, C. W., Chen, C. H., Yu, Y. H., & Tsai, C. J. (2006). Visual sensitivity guided bit allocation for video coding. *IEEE Transactions on Multimedia*, *8*(1), 11–18.
Ugur, K., Andersson, K., Fuldseth, A., Bjontegaard, G., Endresen, L. P., Lainema, J., et al. (2010). High performance, low complexity video coding and the emerging HEVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *20*(12), 1688–1697.
Wang, S., Ma, S., Zhao, D., & Gao, W. (2014). Lagrange multiplier based perceptual optimization for high efficiency video coding. In *Asia-Pacific signal and information processing association, 2014 annual summit and conference (APSIPA)* (pp. 1–4). IEEE.
Wang, S., Rehman, A., Wang, Z., Ma, S., & Gao, W. (2012). SSIM-motivated rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(4), 516–529.
Wang, S., Rehman, A., Wang, Z., Ma, S., & Gao, W. (2013). Perceptual video coding based on SSIM-inspired divisive normalization. *IEEE Transactions on Image Processing*, *22*(4), 1418–1429.
Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.
Wang, Z., Zeng, H., Chen, J., & Cai, C. (2014). Key techniques of high efficiency video coding standard and its extension. In *2014 IEEE 9th conference on industrial electronics and applications (ICIEA)* (pp. 1169–1173). IEEE.
Xu, L., Ma, L., Ngan, K. N., Lin, W., & Weng, Y. (2013). Visual quality metric for perceptual video coding. In *Visual communications and image processing (VCIP)* (pp. 1–5).
Xu, M., Deng, X., Li, S., & Wang, Z. (2014). Region-of-interest based conversational HEVC coding with hierarchical perception model of face. *IEEE Journal of Selected Topics in Signal Processing*, *8*(3), 475–489.
Xue, W., Zhang, L., Mou, X., & Bovik, A. (2014). Gradient magnitude similarity deviation: A highly efficiency perceptual image quality index. *IEEE Transactions on Image Processing*, *23*(2), 684–695.
Yeo, C., Tan, H. L., & Tan, Y. H. (2013). On rate distortion optimization using SSIM. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(7), 1170–1181.
Zeng, H., Ngan, K. N., & Wang, M. (2013). Perceptual adaptive Lagrangian multiplier for high efficiency video coding. In *Picture coding symposium (PCS)* (pp. 69–72). IEEE.
Zeng, H., Yang, A., Ngan, K. N., & Wang, M. (2015). Perceptual sensitivity-based rate control method for high efficiency video coding. In *Multimedia tools and applications* (pp. 1–14).
Zhang, F., Ma, L., Li, S., & Ngan, K. N. (2011). Practical image quality metric applied to image coding. *IEEE Transactions on Multimedia*, *13*(4), 615–624.
Zhao, H., Xie, W., Zhang, Y., Yu, L., & Men, A. (2013). An SSIM-motivated LCU-level rate control algorithm for HEVC. In *Picture coding symposium (PCS)* (pp. 85–88). IEEE.

Aisheng Yang received the B.S. degree from Jiangxi Science & Technology Normal University, Nanchang, China. He is pursuing master degree with the School of Information Science and Engineering, Huaqiao University, Xiamen, China. His research interests are image processing, perceptual video coding, and rate control.

Huanqiang Zeng received the B.S. and M.S. degrees from Huaqiao University, Xiamen, China, and the Ph.D. Degree from Nanyang Technological University, Singapore, all in electrical engineering. He was a Research Associate with Temasek Laboratories, Nanyang Technological University, and a Post-Doctoral Fellow with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with the School of Information Science and Engineering, Huaqiao University, Xiamen China. He has been serving as the Guest/Associate Editor for several international journals, the Area Chair and Technical Program Committee Member for multiple international conferences and a reviewer for multiple international journals and conferences. His research interests include visual information processing, video communication, 3-D/multiview video processing, and computer vision.

Jing Chen received the B.S. and M.S. degrees from Huaqiao University, Xiamen, China, and the Ph.D. degree from Xiamen University, Xiamen, China, all in computer science. She is now an Associate Professor at the School of Information Science and Engineering, Huaqiao University, Xiamen, China. Her current research interests include image processing, multi-view video coding, and multiple description coding.

Jianqing Zhu received the B.S. and M.S. degrees from Huaqiao University, Xiamen, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, China. He is currently an assistant professor with the School of Engineering, Huaqiao University. His research interests are image video processing, face detection, and deep learning.

Canhui Cai received the B.S., M.S., and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, Shanghai University, Shanghai, China, and Tianjin University, Tianjin, China, in 1982, 1985, and 2003, respectively. He has been with the Faculty of Huaqiao University, Quanzhou, China, since 1984, He was a Visiting Professor with Delft University of Technology, Delft, The Netherlands, from 1992 to 1993, and a Visiting Professor with University of California at Santa Barbara, Santa Barbara, CA, USA, from 1999 to 2000. He is currently a Professor with the School of Information Science and Engineering, Huaqiao University. He has authored or co-authored four books and published more than 120 papers in journals and conference proceedings. His research interests include video communications, image and video signal processing, and computer vision. Dr. Cai was a General Co-Chair of Intelligent Signal Processing and Communication Systems in 2007. He is an IEEE senior member.