Closed-Form Optimization on Saliency-Guided Image Compression for HEVC-MSP

Shengxi Li, *Student Member, IEEE*, Mai Xu, *Senior Member, IEEE*, Yun Ren, *Student Member, IEEE*, and Zulin Wang, *Member, IEEE*

***Abstract*—High efficiency video coding (HEVC) is the latest video coding standard, and it has the best performance among all the existing standards. HEVC main still picture profile (HEVC-MSP) also achieves top performance in image compression. In this paper, we propose a closed-form bit allocation approach to optimize the saliency-guided PSNR (viewed as perceptual distortion) such that the coding efficiency of HEVC-based image compression can be significantly improved from a subjective perspective. Specifically, a bit allocation formulation is established to minimize perceptual distortion with a constraint on bit-rates. Then, this formulation is solved using the proposed recursive Taylor expansion method with a closed-form solution. On the basis of our solution, a bit allocation and re-allocation process is developed in our approach to minimize perceptual distortion, meanwhile accurately controlling bit-rates. In addition, we provide both theoretical and numerical analyses of the computational complexity, verifying the little extra time cost of our approach. The experimental results demonstrate the superior performance of our approach over the state-of-the-art HEVC-MSP, and the BD-rate savings are approximately 40% and 24% for face and generic images, respectively.**

***Index Terms*—High efficiency video coding (HEVC), perceptual image compression, saliency detection.**

I. Introduction

At PRESENT, multimedia applications, such as Facebook and Twitter, are becoming integral components in the daily lives of millions, leading to the explosion of big data. Among them, images are one of the largest types of big data [2], thus posing a great challenge to the limited communication and storage resources. As reported by [3], more than one million images are “making their way” to Facebook every hour. Meanwhile, due to more powerful camera hardware, the resolutions of images are significantly increasing, further intensifying the hunger on communication and storage resources. Aiming at overcoming this resource-hungry issue, a set of image compression standards have been proposed to condense image data, e.g., JPEG [4], JPEG 2000 [5], JPEG XR [6], and WebP [7]. Recently, some cloud-based image compression methods (e.g., [8]) have also provided a promising way to compress one image using a number of similar images in the cloud.

Compared with image compression standards, several video coding standards, such as H.264/AVC [9] and VP9 [10], have shown the same or even better performance for compressing still images. Most recently, as the successor of H.264/AVC, High Efficiency Video Coding (HEVC) [11] was formally approved in April, 2013. In HEVC, several new features, e.g., the quadtree-based coding structure and intra prediction modes with 33 directions,[1] were adopted. Consequently, the HEVC Main Still picture (HEVC-MSP) profile [12], which is designed for still picture compression, achieves the best performance among all the state-of-the-art standards on image compression, with an approximately 10% (over VP9) - 40% (over JPEG) improvement in bit-rate savings [13]. However, all existing standards, including HEVC-MSP, primarily focus on removing statistical redundancy by adopting various techniques [14], e.g., intra prediction and entropy coding. Further reducing statistical redundancy may help to improve coding efficiency, but at the cost of extremely high computational complexity.

Koch *et al.* [15] investigated that the bandwidth between the human eyes and brain is approximately 8 Mbps, which is far insufficient to process the visual input captured by millions of optical cells. Thus, the human eye is mostly at a quite low resolution, except for a small area at the fovea (visual angle of approximately 2°), which is called the region-of-interest (ROI) in the image compression community. Meanwhile, as pointed out by [16], human ROIs are similar across different individuals. It is also well known [17] that the coding mechanism can be modified to cater to the HVS by moving bits from non-ROIs to ROIs to achieve better subjective quality. This is also illustrated in Fig. 1(b) and 1(c). Several perceptual image compression approaches [18]–[21] have been developed on the basis of this modification. For example, in [18], the diagnostically useful regions (i.e., ROIs) are encoded losslessly using S-transform, whereas the other regions are compressed using a lossy wavelet

Manuscript received March 1, 2016; revised March 29, 2017 and June 12, 2017; accepted June 23, 2017. Date of publication June 29, 2017; date of current version December 14, 2017. This work was supported by the National Nature Science Foundation of China under Grant 61573037 and by the Fok Ying Tung Education Foundation under Grant 151061. This paper was presented at the Data Compression Conference, Snowbird, UT, USA, March-April 2016 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Balakrishnan Prabhakaran. *(Corresponding author: Mai Xu.)*

S. Li, M. Xu, and Y. Ren are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: ShengxiLi2014@gmail.com; MaiXu@buaa.edu.cn; Yunren@buaa.edu.cn).

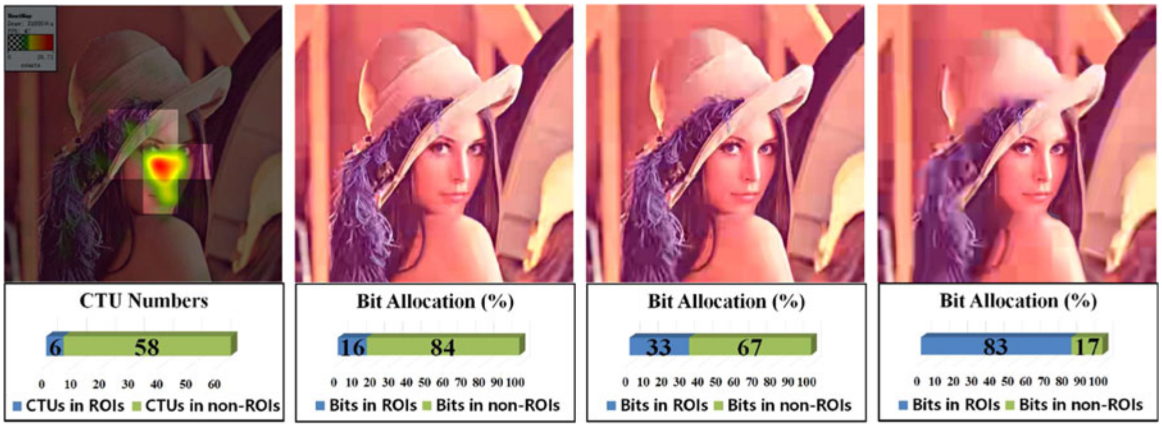
Z. Wang is with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430072, China, and also with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: wzulin@buaa.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

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/TMM.2017.2721544

[1] Planar and DC are two other intra prediction modes.



(a) Heat map (b) No emphasis (c) Well balanced (d) More emphasis

Fig. 1. Example of HEVC-based image compression for the *Lena* image, with different bit allocation emphasis on ROIs. Note that (a) is the heat map of eye fixations; (b), (c), and (d) are compressed by HEVC-MSP at 0.1 bpp with no, well-balanced, and more emphasis on face regions. The DMOS scores (to be discussed in Section V) for (b), (c), and (d) are 63.9, 57.5, and 70.3, respectively. (a) Heat map. (b) No emphasis. (c) Well-balanced. (d) More emphasis.

zerotree. In this way, the ROIs can be ensured with high quality to improve the overall subjective quality. For generic images, an ROI-based set partitioning in hierarchical trees (SPIHT) algorithm was proposed in [19]. In this algorithm, the ROI compression is achieved by modifying the information order of the SPIHT structure and by emphasizing the transform coefficients belonging to the ROIs. In [20], perceptual image compression is achieved by maintaining the discrete wavelet transform (DWT) coefficients in ROIs while reducing some coefficients in non-ROIs, which is based on a type of saliency detection method. In addition, in [21], the ROIs specified by the users are endowed with high priority by locating the corresponding segments in the master bitstream of JPEG 2000. However, through our investigation, the substantial low quality in non-ROIs may also significantly degrade image quality, as shown in Fig. 1(d). Thus, how many bits should "move" from non-ROIs to ROIs, together with accurate ROI detection, is crucial for image compression. In other words, we need to ensure that the detected ROIs are the regions that attract human attention, and then bit allocation needs to be optimized according to ROIs, targeting minimal overall perceptual distortion. To our best knowledge, there exists no bit allocation work that has a closed-form optimization on the subjective quality of compressed images with a bit-rate constraint, in the state-of-the-art HEVC-based image compression.

In this paper, we propose a closed-form bit allocation approach to minimize the perceptual distortion. Consequently, the coding efficiency of the state-of-the-art HEVC-MSP can be significantly improved from a subjective perspective. Specifically, the most recent work [22] has pointed out that eye-tracking weighted PSNR (EWPSNR), which is the combination of eye-tracking fixations and mean square error (MSE), is highly correlated with subjective quality. Due to the unavailability of eye-tracking data, we utilize the saliency weighted PSNR (SW-PSNR) instead as the perceptual distortion to approximate subjective quality. Automatic saliency detection is thus the first step of our approach for saliency-guided image compression. In our approach, we leverage on our most recent face saliency detection method [23] for compressing face images and a latest saliency detection method [24] for compressing other generic images. Note that face and non-face images are automatically classified using the face detector in [23]. Then, we propose a formulation to minimize perceptual distortion with reasonable bit allocation on compressed images. Unfortunately, it is intractable to obtain a closed-form solution to the proposed optimization formulation because the formulation is a high-order algebraic equation, and its non-integer exponents vary across different coding tree units (CTUs). We thus develop a new method, namely, recursive Taylor expansion (RTE), to acquire the solution for optimal bit allocation in a closed-form manner. In the proposed RTE method, we iterate a third-order Taylor expansion to reach the optimal solution for bit allocation. We also develop an optimal bit reallocation process to alleviate the mismatch between the target and actual bits, while maintaining perceptual distortion optimization. We further verify via both theoretical and numerical analyses that little time cost is incurred by our approach.

This paper is an extended version of our conference paper [1] with extensive advancements. Specifically, the application of [1] only focuses on face images, whereas this paper extends the application to all generic images by adopting a saliency detection method [24] for non-face image compression. This paper also advances the derivation of the proposed RTE method with more thorough analysis and solid proofs. Beyond the numerical analysis of [1], the theoretical analysis on computational complexity is also provided in this paper. Additionally, more comprehensive experimental results are presented to validate the rate-distortion (R-D) performance of our approach. In addition to the R-D evaluation, this paper also enhances the assessment by offering the accuracy of rate control (RC) at both the picture and CTU levels. In summary, our main contributions are as follows:

1) We present a formulation to optimize the perceptual distortion of HEVC-based image compression by adopting image saliency in the subjective quality metric.

2) We propose the RTE method to solve the optimization formulation on bit allocation, followed by a bit re-allocation process to accurately control bit-rates.
3) We analyze the computational complexity using both theoretical and numerical aspects, verifying the little extra time cost.

The contribution of our paper primarily differs from the state-of-the-art in three aspects. First, our approach designs a formulation on subjective quality optimization for the latest image compression standard of HEVC-MSP, whereas the majority of other works focus on the optimization for previous image compression standards, such as JPEG, JPEG 2000, and JPEG-XR [18]–[20], [25]–[51]. Second, our approach derives the closed-form solution with little extra time on optimizing the subjective quality for image compression. In contrast, the existing works do not include any optimization[2] [18]–[20], [34]–[41] or have sub-optimal solutions [42]–[51]. Third, our approach involves an optimal bit re-allocation process for accurate RC, whereas the state-of-the-art RC approaches [52], [53] in HEVC-MSP equally re-allocate bits or even do not incorporate any bit re-allocation process [18]–[20], [25]–[51].

The remainder of this paper is organized as follows. In Section II, we review the related works on perceptual image compression. Then, Section III provides the details of the proposed approach. Subsequently, the computational complexity of the proposed approach is analyzed in Section IV. Finally, Section V presents the experimental results, and Section VI concludes this paper.

II. Related Works

The main goal of image compression is to enhance coding efficiency by either reducing bits or improving quality. More importantly, the quality perceived by the HVS, called subjective quality, needs to be improved for enhancing coding efficiency since human beings are the final ends of image compression. However, MSE, a common metric of visual quality, has been argued [54] in many works to be insufficient in terms of correlation with subjective quality. Consequently, extensive approaches [55] focus on exploiting the HVS to improve subjective quality rather than MSE, which fall in the scope of perceptual image compression.

Many ongoing approaches on mimicking the HVS have shed some light on perceptual image compression. Several common features of the HVS [55], [56] have been studied and then applied in image compression. In terms of enhancing the coding efficiency of perceptual image compression, the related works can mainly be classified into two categories: bit reduction to maintain a desired perceptual distortion and quality improvement with a bit-rate constraint.

A. Bit Reduction at a Desired Perceptual Distortion

Many approaches [25]–[33] incorporate just noticeable difference (JND) [57] and other features of the HVS to save bit-rates while maintaining an almost unchanged subjective quality for the compressed images. JND represents the discrimination ability on the difference between two or more stimuli. Based on JND, the just not noticeable difference (JNND) threshold is utilized in [30] for the perceptual compression of medical images to reduce the irrelevant information. Recently, based on the free-energy principle, an advanced JND model was proposed in [31]. With this JND model, the bit-rates can be saved for image compression. Moreover, in [32], an adaptive down-sampling coder was proposed to save bits and computational complexity by comparing whether the differences between down-sampled and original pixels exceed a pixel-wise JND model. Furthermore, to minimize bits at a given perceptual distortion, a discrete cosine transform (DCT)-based locally adaptive perceptual image compression [29] was proposed to iteratively approach the desired perceptual distortion, in which the JND threshold is estimated via contrast sensitivity on background luminance and contrast masking. In addition to contrast sensitivity and contrast masking, Liu *et al.* [27] adopted an additional factor to calculate JND, i.e., luminance masking, for perceptual compression in JPEG 2000. Similar to [29], by iterating to reach the desired distortion, the minimum bits can be achieved in [27]. Recently, Zhang *et al.* [33] proposed to optimize the overall rate-distortion performance across all DCT bands according to a derived JND-based quantization table. Then, it iterates to reach the target distortion while maintaining the minimum bits. However, the above approaches are too time consuming to be applied due to the brute force search for the optimal solution.

In addition to JND models, other bit reduction approaches have also been developed for perceptual image compression, e.g., the suprathreshold distortion approaches [58], edge-based approaches [59], and other bio-inspired approaches [60]. Nevertheless, all these approaches for bit reduction can hardly be used in resource-limited applications, in which the subjective quality needs to be improved at given bit-rates.

B. Subjective Quality Improvement With a Constraint on Bit-Rate

When the bandwidth and storage resources are limited, people prefer to "receive" subjective quality that is as favorable as possible. The approaches for subjective quality improvement thus provide the accessibility to this end [18]–[20], [34]–[41]. A common way to achieve this goal in these approaches is to allocate relatively more bits in ROIs to ensure acceptable quality in these regions. For medical images, ROI-based image compression has been widely studied [35]. Later, Liu *et al.* [34] proposed a significant bitplanes shift (PSBShift) approach to ensure higher quality in ROIs than in non-ROIs for perceptual JPEG 2000, which flexibly combines two types of ROI-based methods [61], [62]. Recently, an advanced PSBShift approach was proposed in [39] for more flexible RC in JPEG 2000. Moreover, with the progressive streaming of JPEG 2000 and JPEG, a novel ROI image compression scheme [38] was proposed to improve the quality of ROIs. This scheme adopts the rate-distortion-complexity tradeoff with a jointly suboptimized residual vector quantizer (JSRVQ) method. Moreover, in [40], the less blurred regions are considered to be the ROIs, which are allocated with

[2] Those approaches only increase the amount of bits in ROIs.

more bits. Besides, benefiting from the most recent deep learning technique, the ROIs are automatically detected in [41] by a convolutional neural network (CNN) and are then encoded with higher quality. The above approaches primarily improve the fidelity of ROIs, but they may fail in ensuring the overall subjective quality, as extremely low quality on non-ROIs can also degrade the subjective quality. In this paper, our approach optimizes the overall subjective quality for HEVC-MSP, different from the above approaches that only increase bits in ROIs.

Several approaches have been proposed to improve the overall subjective quality [42]–[51]. The initial approach of exploiting rate-versus-distortion can be traced back to [42] for monochrome images. In [42], the perceptual distortion, modeled by weighted MSE (WMSE), is minimized by an empirical optimal weighting function. Recently, Chen *et al.* [45] proposed to automatically produce probable ROI masks with a specified initial point in JPEG 2000. By embedding such masks into rate-WMSE optimization, the images can be compressed with favorable perceptual quality and high PSNR values. In addition, Channappayya *et al.* [44] adopted structural similarity (SSIM) for perceptual distortion optimization. In their work, the optimal bit allocation is approached by a bound constraint mechanism on SSIM in the DCT domain, thus realizing optimization on SSIM. Later, a multi-scale mean SSIM (MSSIM) was applied in [48] as the metric for estimating the overall subjective quality. Then, the bit allocation was optimized by iterating the quantization parameters (QPs) to make the multi-scale MSSIM of each block roughly identical. Unfortunately, it is intractable to establish a closed-form relationship between bit-rates and SSIM [44], thus leading to sub-optimal results during bit allocation. Furthermore, in [50], an SSIM-based metric was utilized and optimized to choose the best coding tiling from a multi-tree dictionary. As the optimal result cannot be analytically solved, an MSE-based RDO was adopted as an alternative in [50], followed by a dynamic programming technique to find the optimal tiling mode. Moreover, a perceptual compression work towards high dynamic range images was also proposed in [51], which introduces an iteration process to optimize the enhanced tone mapped image quality index (TMQI). In [49], a perceptual coding scheme was proposed by adopting a simple yet effective metric, which combines the texture masking effect and contrast sensitivity function. However, this scheme requires a post-processing on the decoder side. To effectively bridge the gap between perception and bit-rates, another way is to take into account visual attention in image compression, the effectiveness of which has been verified in [63]. In this paper, our approach can obtain the closed-form solution with little extra time for optimizing the overall quality, rather than the sub-optimization solution in the above works.

Specifically, we propose a closed-form solution for the bit allocation of HEVC-based image compression, which optimizes the perceptual distortion. The perceptual distortion is measured in terms of the combination of saliency and MSE. Our motivations are three-fold: 1) HEVC-MSP retains the top performance among all existing standards [13], 2) combining saliency weight and MSE is simple yet effective in modeling subjective quality [63], and 3) our approach enjoys the closed-form bit allocation (by our RTE method) to minimize perceptual distortion for HEVC-MSP. Consequently, the coding efficiency of HEVC-MSP can be greatly enhanced at a given bit-rate from the perspective of subjective quality.

III. Minimizing Perceptual Distortion With Our RTE Method

In this section, we primarily focus on minimizing the perceptual distortion of HEVC-MSP, i.e., catering to the visual quality of detected ROIs. To this end, we first transplant the R-λ RC approach [53] into HEVC-MSP in Section III-A. Upon this, an optimization formulation is proposed in Section III-B, which aims at maximizing the SWPSNR at a given bit-rate for each image. The RTE method is then proposed in Section III-C to solve this formulation with a closed-form solution. In this way, the perceptual distortion can be minimized via bit allocation. In addition, we develop an optimal bit re-allocation method in Section III-D to alleviate the mismatch between the target and actual bit-rates.

A. Rate Control Implementation on HEVC-MSP

The latest R-λ approach is proposed in [53] for RC in HEVC. Since we concentrate on applying RC to image compression, the CTU level RC in one video frame is discussed here. Specifically, for HEVC, it has been verified that the hyperbolic model can better fit the rate-distortion (R-D) relationship [53]. Based on the hyperbolic model, an R-λ model is developed for bit allocation in the latest HEVC RC approach, where λ is the slope of the R-D relationship [64]. Assuming that d_i, r_i and λ_i represent the distortion, bits and R-D slope for the i-th CTU, respectively, the R-D relationship and R-λ model are formulated as follows:

$$d_i = c_i {r_i}^{-k_i} \tag{1}$$

and

$$\lambda_i = -\frac{\partial d_i}{\partial r_i} = c_i k_i \cdot {r_i}^{-k_i-1} \tag{2}$$

where c_i and k_i are the parameters that reflect the content of the i-th CTU. In the R-λ approach [53], r_i is first allocated according to the predicted mean absolute difference (MAD), and then its corresponding λ_i is obtained using (2). By adopting a fitting relationship between λ_i and QP, the QPs of all CTUs within the frame can be estimated such that RC is achieved in HEVC. For more details, refer to [53].

However, for HEVC-MSP, c_i and k_i cannot be obtained when encoding CTUs. Thus, it is difficult to directly apply the R-λ RC approach to HEVC-MSP. In the work of [52], the sum of the absolute transformed differences (SATD), calculated by the sum of Hadamard transform coefficients, is utilized for HEVC-MSP. Specifically, the modified R-λ model is

$$\lambda_i = \alpha_i \left(\frac{s_i}{r_i}\right)^{\beta_i} \tag{3}$$

where α_i and β_i are constants for all CTUs and remain the same when encoding an image. Moreover, s_i denotes the SATD for the i-th CTU, which measures the CTU texture complexity.

Nevertheless, SATD is too simple to reflect image content, leading to an inaccurate R-D relationship during RC.

To avoid the above issues, we adopt a pre-processing process in calculating c_i and k_i. After pre-compressing, the pre-encoded distortion, bits and λ can be obtained for the i-th CTU, which are denoted as $\bar{d}_i$, $\bar{r}_i$ and $\bar{\lambda}_i$, respectively. Then, the RC-related parameters, c_i and k_i, can be estimated upon (1) and (2) before encoding the i-th CTU

$$c_i = \frac{\bar{d}_i}{\left(\bar{r}_i^{-\bar{\lambda}_i \cdot \bar{r}_i / \bar{d}_i}\right)} \tag{4}$$

and

$$k_i = \frac{\bar{\lambda}_i \cdot \bar{r}_i}{\bar{d}_i}. \tag{5}$$

With the estimated c_i and k_i, the RC of the R-λ approach [53] can be implemented in HEVC-MSP.

Here, a fast pre-compressing process is developed in our approach, which sets the maximum coding unit (CU) depth to 0 for all CTUs. We have verified that the fast pre-compressing process slightly increases the computational complexity by a 5% burden, which is slightly larger than the 3% of the SATD-based method [52]. However, this process is able to well reflect the R-D relationship, as to be verified Section V-D.

B. Optimization Formulation on Perceptual Distortion

The primary objective of this paper is to maximize perceptual distortion for HEVC-based image compression. In our approach, the SWPSNR is applied to measure the perceptual distortion, as [63] has shown that SWPSNR is highly correlated with subjective quality. For SWPSNR, the pixel-wise saliency values need to be detected as the first step in our approach, and these values are used for weighting the MSE. In this paper, we utilize two state-of-the-art saliency detection methods for calculating SWPSNR. Specifically, the latest boolean map based saliency (BMS) method [24] is applied in modeling SWPSNR for generic images. Furthermore, for face images, our most recent work [23] has better accuracy in saliency detection than the BMS method. Thus, when computing the SWPSNR of face images, we use the work of [23] to obtain the saliency values.

Here, we denote w_i as the average saliency value within the i-th CTU. Meanwhile, we calculate distortion d_i by the sum of pixel-wise square error for the i-th CTU. Then, based on d_i and w_i, the optimization on SWPSNR at a given target bit-rate R can be formulated as

$$\min \left(\frac{\Sigma_{i=1}^{M} w_i d_i}{\Sigma_{i=1}^{M} w_i}\right) \quad \text{s.t. } \Sigma_{i=1}^{M} r_i = R. \tag{6}$$

In (6), M denotes the number of CTUs in the image. By using the Lagrange multiplier λ, (6) can be turned to find the minimum value of R-D cost J [64], which is defined as

$$J = \left(\frac{\Sigma_{i=1}^{M} w_i d_i}{\Sigma_{i=1}^{M} w_i}\right) + \lambda \cdot (\Sigma_{i=1}^{M} r_i). \tag{7}$$

By setting the partial derivatives of (7) to zero, the minimum J can be found as follows:

$$\begin{aligned} \frac{\partial J}{\partial r_i} &= \frac{\partial \left(\Sigma_{i=1}^{M} w_i d_i / \Sigma_{i=1}^{M} w_i + \lambda(\Sigma_{i=1}^{M} r_i)\right)}{\partial r_i} \\ &= \frac{w_i}{\Sigma_{i=1}^{M} w_i} \cdot \frac{\partial d_i}{\partial r_i} + \lambda \\ &= 0. \end{aligned} \tag{8}$$

Given (1) and (2), (8) is turned to

$$r_i = \left(\frac{\lambda \cdot \Sigma_{i=1}^{M} w_i}{c_i k_i w_i}\right)^{-\frac{1}{k_i+1}} = \left(\frac{\widetilde{w}_i a_i}{\lambda}\right)^{b_i} \tag{9}$$

where $a_i = c_i k_i$ and $b_i = \frac{1}{k_i+1}$ also reflect the image content for each CTU. Moreover, $\widetilde{w}_i = w_i / (\Sigma_{i=1}^{M} w_i)$ represents the visual importance for each CTU. Note that with our pre-compressing process, c_i and k_i can be obtained in advance. Thus, a_i and b_i are available before encoding the image. Once λ is known, r_i can be estimated using (9) for achieving the minimum J.

Meanwhile, there also exists a constraint on bit-rate, which is formulated as

$$\sum_{i=1}^{M} r_i = R. \tag{10}$$

According to (9) and (10), we need to find the "proper" λ and bit allocation r_i to satisfy the following equation:

$$\sum_{i=1}^{M} r_i = \sum_{i=1}^{M} \left(\frac{\widetilde{w}_i a_i}{\lambda}\right)^{b_i} = R. \tag{11}$$

After solving (11) to find the "proper" λ, the target bits can be assigned to each CTU with the maximum SWPSNR.

Unfortunately, since a_i and b_i vary across different CTUs, (11) cannot be solved by a closed-form solution. Next, the RTE method is proposed to provide a closed-form solution.

C. RTE Method for Solving the Optimization Formulation

For solving (11), we assume that $\widetilde{r}_i(\widetilde{\lambda})^{b_i} = (\widetilde{w}_i a_i)^{b_i}$, where $\widetilde{r}_i$ and $\widetilde{\lambda}$ are the estimated r_i and λ, respectively. Then, (11) can be rewritten as

$$\sum_{i=1}^{M} r_i = \sum_{i=1}^{M} \left(\frac{\widetilde{w}_i a_i}{\lambda}\right)^{b_i} = \sum_{i=1}^{M} \widetilde{r}_i \left(\frac{\widetilde{\lambda}}{\lambda}\right)^{b_i} = R. \tag{12}$$

From (12), we can see that once $\widetilde{\lambda} \rightarrow \lambda$, there exists $\widetilde{r}_i \rightarrow r_i$. As such, the optimization formulation of (11) can be solved in our approach. However, we do not know $\widetilde{\lambda}$ at the beginning. Meanwhile, λ of (12) is also unknown because it is intractable to find the closed-form solution to (11). Therefore, a chicken-and-egg dilemma exists between $\widetilde{\lambda}$ and λ. To solve this dilemma, a possible $\widetilde{\lambda}$ is initially set. In our RTE method, the picture λ (denoted as λ_{pic}) is chosen as the initial value of $\widetilde{\lambda}$ for quick convergence. It is calculated by the R-λ model at the picture level [52], [53]

$$\lambda_{pic} = \alpha_{pic} \left(\frac{s_{pic}}{R}\right)^{\beta_{pic}} \tag{13}$$

where α_{pic} and β_{pic} are the fitted constants ($\alpha_{pic} = 6.7542$ and $\beta_{pic} = 1.7860$ in HM 16.0); s_{pic} represents the SATD for the current picture. Recall that R denotes the target bits allocated to the currently encoded picture.

In the following, the RTE method is proposed to iteratively update $\widetilde{\lambda}$ for making $\widetilde{\lambda} \to \lambda$.

Specifically, we preliminarily apply Taylor expansion on $(\frac{\widetilde{\lambda}}{\lambda})^{b_i}$ of (12), and then we discard the biquadratic and higher-order terms. The process can be formulated as follows:

$$\begin{aligned} R &= \sum_{i=1}^{M} \widetilde{r}_i \left(\frac{\widetilde{\lambda}}{\lambda}\right)^{b_i} \\ &= \sum_{i=1}^{M} \widetilde{r}_i \left(1 + \frac{\ln(\frac{\widetilde{\lambda}}{\lambda})}{1!} b_i + \cdots + \frac{(\ln\frac{\widetilde{\lambda}}{\lambda})^n}{n!} b_i^n + \cdots\right) \\ &\approx \sum_{i=1}^{M} \widetilde{r}_i \left(1 + \frac{\ln(\frac{\widetilde{\lambda}}{\lambda})}{1!} b_i + \frac{(\ln\frac{\widetilde{\lambda}}{\lambda})^2}{2!} b_i^2 + \frac{(\ln\frac{\widetilde{\lambda}}{\lambda})^3}{3!} b_i^3\right). \end{aligned} \tag{14}$$

Here, we use $\widehat{\lambda}$ to denote the approximation solution to (14) after discarding the biquadratic and higher-order terms. Consequently, (12) can be approximated to be a cubic equation with variable $\ln \widehat{\lambda}$

$$\begin{aligned} R = \sum_{i=1}^{M} \widetilde{r}_i & \left(1 + \frac{\ln(\frac{\widetilde{\lambda}}{\widehat{\lambda}})}{1!} b_i + \frac{(\ln\frac{\widetilde{\lambda}}{\widehat{\lambda}})^2}{2!} b_i^2 + \frac{(\ln\frac{\widetilde{\lambda}}{\widehat{\lambda}})^3}{3!} b_i^3\right) \\ = \quad & \underbrace{-\sum_{i=1}^{M} \widetilde{r}_i (\frac{b_i^3}{6})}_{A} \ln^3\widehat{\lambda} + \underbrace{\sum_{i=1}^{M} \widetilde{r}_i (\frac{b_i^2}{2} + \frac{b_i^3}{2} \ln\widetilde{\lambda})}_{B} \ln^2\widehat{\lambda} \\ & \underbrace{-\sum_{i=1}^{M} \widetilde{r}_i (b_i^2 \ln\widetilde{\lambda} + b_i + \frac{b_i^3}{2} \ln^2\widetilde{\lambda})}_{C} \ln\widehat{\lambda} \\ & \underbrace{+\sum_{i=1}^{M} \widetilde{r}_i (1 + b_i \ln\widetilde{\lambda} + \frac{b_i^2}{2} \ln^2 \widetilde{\lambda} + \frac{b_i^3}{6} \ln^3 \widetilde{\lambda})}_{D}. \end{aligned} \tag{15}$$

By applying the Shengjin formula [65], this cubic equation is evaluated to obtain the solution of $\widehat{\lambda}$ as

$$\widehat{\lambda} = e^{\frac{-B - (\sqrt[3]{Y_1} + \sqrt[3]{Y_2})}{3A}}, Y_{1,2} = BE + 3A\left(\frac{-F \pm \sqrt{F^2 - 4EG}}{2}\right) \tag{16}$$

where $E = B^2 - 3AC$, $F = BC - 9A(D - R)$, and $G = C^2 - 3B(D - R)$. Since $\Delta = F^2 - 4EG > 0$ in practical encoding, (16) has only one real solution [65]. Thus, the value of $\widehat{\lambda}$ is unique for optimizing bit allocation. After further removing the cubic-order term, (14) is turned to be a quadratic equation. We found that such a quadratic equation may have no real solution or two solutions. Meanwhile, using only one term may lead to large approximation error and slow convergence speed, whilst keeping more than four terms probably makes the polynomial equations on $\ln \widehat{\lambda}$ unsolvable.

TABLE I
RTE Method for Solving (12)

```
–  Input: $a_i$, $b_i$, $w_i$ for each encoding CTUs and target bits R.
–  Output: reasonable bit allocation $\widetilde{r}_i$ for each CTU on maximizing
   SWPSNR.
•  Initialize $\widetilde{\lambda}$ to be $\lambda_{pic}$.
•  While $\widetilde{\lambda}$ does not meet the convergence criterion
       1  Calculate $A$, $B$, $C$, and $D$ of (15) with $\widetilde{\lambda}$.
       2  Obtain $\widehat{\lambda}$ estimated by (16).
       3  Update $\widetilde{\lambda}$ with the obtained $\widehat{\lambda}$.
•  End
•  Save the final $\widetilde{\lambda}$.
•  Apply it to bit allocation $\widetilde{r}_i$ with (9)
•  Return $\widetilde{r}_i$ for each CTU.
```

Therefore, discarding the biquadratic and higher-order terms of the Taylor expansion is the best choice for our approach.

However, due to the truncation of high-order terms in the Taylor expansion, $\widehat{\lambda}$ estimated by (16) may not be an accurate solution to (12). Fortunately, as proven in Lemma 1, $\widehat{\lambda}$ is more accurate[3] than $\widetilde{\lambda}$ when $\widetilde{\lambda} < \lambda$.

Lemma 1: Consider $\lambda > \widetilde{\lambda} > 0$, $b_i > 0$, and $R > 0$ for (12). When the solution of λ to (12) is $\widehat{\lambda}$, the following inequality holds for $\widehat{\lambda}$

$$|\widehat{\lambda} - \lambda| < |\widetilde{\lambda} - \lambda|. \tag{17}$$

Proof: Please refer to the supporting document at https://github.com/RenYun2016/TMM2016.

As shown in Lemma 1, although both $\widetilde{\lambda}$ and $\widehat{\lambda}$ may be inaccurate for estimating λ in (11), $\widehat{\lambda}$, obtained through (12)–(16), is closer to λ than $\widetilde{\lambda}$. Therefore, we can iterate the Taylor expansion by using $\widehat{\lambda}$ as $\widetilde{\lambda}$ to the next iteration, which is the core of our RTE method. In addition, if $\widetilde{\lambda} > \lambda > 0$ at the first iteration, then its output $\widehat{\lambda}$ is smaller than λ, as pointed out by Lemma 2.

Lemma 2: Consider $\widetilde{\lambda} > 0, \lambda > 0, b_i > 0, \lambda \neq \widetilde{\lambda}$, and $R > 0$ for (12). If $\widehat{\lambda}$ is the solution of λ to (12), then the following holds:

$$\widehat{\lambda} < \lambda. \tag{18}$$

Proof: Please refer to the supporting document at https://github.com/RenYun2016/TMM2016.

Given Lemma 2, for the subsequent iterations of the RTE method, $0 < \widetilde{\lambda} < \lambda$ of Lemma 1 can be satisfied since the value of $\widetilde{\lambda}$ has been replaced by that of $\widehat{\lambda}$. In this way, the closed-form solution λ can be obtained by iteratively estimating $\widetilde{\lambda}$.

Our RTE method is summarized in Table I. For each iteration, the convergence criterion is set according to the approximation error, $E_a < 10^{-10}$, where $E_a = |\Sigma_{i=1}^{M} \widetilde{r}_i - R|/R$. As analyzed in Section IV, the approximation error of our RTE method is able to converge to 10^{-10}, generally with no more than three iterations. In other words, after three or fewer iterations, the RTE method is able to reduce the difference between $\widetilde{\lambda}$ and λ to an extremely small range, meeting the convergence criterion. Thus, $\widetilde{\lambda}$ can be output as the closed-form solution to (12) (as well as (11)). Finally, we replace λ by $\widetilde{\lambda}$ in (9) to allocate the target bits to each CTU such that SWPSNR can be maximized.

[3] It is obvious that $0 < b_i = \frac{1}{k_i + 1} < 1$ and $R > 0$ in HEVC encoding.

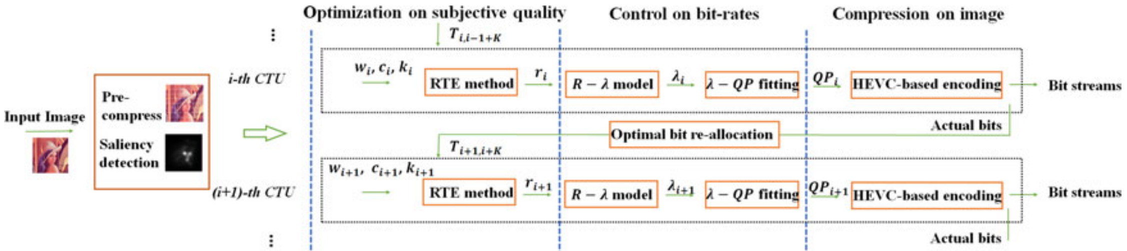


Fig. 2. Procedure of our approach on minimizing perceptual distortion.

The physical explanation for the fast convergence speed of our RTE method is as follows. Obviously, the approximation error for each iteration of the RTE method is largely related to $\ln\frac{\widetilde{\lambda}}{\lambda}$ in $\frac{\widetilde{(\ln\frac{\lambda}{\lambda})^n}}{n!}b_i^n$ of (14). To reduce the value of $\ln\frac{\widetilde{\lambda}}{\lambda}$ for small approximation error, our RTE method utilizes a more accurate solution $\widehat{\lambda}$ after each iteration to replace $\widetilde{\lambda}$ for the next iteration, making $\frac{\widetilde{(\ln\frac{\lambda}{\lambda})^n}}{n!}b_i^n$ decrease sharply. Therefore, such a replacement not only provides a more accurate input for the next iteration but also greatly reduces the values of the discarded terms and the approximation error. In this way, the convergence speed can be accelerated along with iterations. Moreover, keeping three terms for the Taylor expansion rather than other terms is solvable and also contributes to the fast convergence speed of our RTE method.

D. Bit Re-allocation for Maintaining Optimization

As we discussed in Section III-C, bits are reasonably allocated in our approach to minimize perceptual distortion. However, in practical encoding, a slight difference between the target and actual bits may exist for each CTU. This difference may degrade RC accuracy. To overcome this, we develop a bit re-allocation process to accurately control bit-rates, meanwhile maintaining the optimization for perceptual distortion.

Specifically, for compensating the bit-rate error after encoding the i-th CTU, the target bits for the incoming K CTUs (denoted as $T_{i+1,i+K}$) are updated by

$$T_{i+1,i+K}=\sum_{j=i+1}^{j=i+K}\widetilde{r}_j+\underbrace{\left(\widehat{T}-\sum_{j=i+1}^{j=M}\widetilde{r}_j\right)}_{\text{bit}-\text{rate error}}. \tag{19}$$

In (19), $\widehat{T}$ is the remaining bits for encoding remaining CTUs, and $\widetilde{r}_j$ represents the target bits for the j-th CTU by our RTE method. Recall that M denotes the total number of CTUs. Obviously, as seen from (19), the bit error is compensated during encoding the next K CTUs. Here, the RTE method of Section III-C is applied to re-allocate $T_{i+1,i+K}$ to the next K CTUs. Note that we follow [52] and [53] to set $K=4$, which means that bits are re-assigned in the next four CTUs. Moreover, note that due to the fast convergence speed of our RTE method, the complexity increases little for the bit re-allocation process.

Finally, we summarize our HEVC-based image compression approach in Fig. 2. Specifically, we first transplant RC to HEVC-MSP with a simplified pre-compression process, and the saliency values are detected for the input image. Then, our RTE method obtains the target bits of each CTU, which can minimize perceptual distortion at a given bit-rate. Next, the QP value of each CTU is estimated using the R-λ model and QP fitting. Note that the bits need to be re-allocated in the following CTUs to bridge the gap between the target and actual bits. In addition, as to be verified in Section IV, little computational complexity cost is introduced in our RTE method, further highlighting the efficiency of our approach.

IV. Computational Complexity Analysis

In this section, we primarily focus on the computational complexity of our approach. Since our approach adopts the RTE method to optimize perceptual distortion, the convergence speed of the RTE method is first discussed from both theoretical and numerical perspectives. In the numerical analysis, we also provide the practical computational time of our approach.

A. Theoretical Analysis

For the theoretical analysis, we investigate the difference between $\widetilde{\lambda}$ and λ alongside the iterations of our RTE method. Here, we define $\Delta\lambda$ as the difference between $\widetilde{\lambda}$ and λ as

$$\Delta\lambda=\frac{\widetilde{\lambda}-\lambda}{\lambda}. \tag{20}$$

If $|\Delta\lambda|\rightarrow 0$, then it indicates that our RTE method is stably convergent. Therefore, we take into consideration $\Delta\lambda$ along with each iteration in our RTE method to analyze its convergence speed.

In practice, k_i (>0) of (9) varies in a small range when encoding images using HEVC-MSP. Therefore, we assume that b_i ($0<b_i=\frac{1}{k_i+1}<1$ in (9)) remains constant for simplicity. Based on this assumption, the convergence speed of our RTE method can be determined with Lemma 3.

Lemma 3: Consider that $\widetilde{\lambda}>0, \widehat{\lambda}>0, \lambda>0, R>0$, and $\forall i$, $b_i=l\in(0,1)$. Recall that $\widetilde{\lambda}$ is the estimated λ of (12) before each iteration of our RTE method and that $\widehat{\lambda}$ is the solution of λ to (12) after each iteration of our RTE method. After each iteration in our RTE method, $\widetilde{\lambda}$ is replaced by $\widehat{\lambda}$. Then, there exists $|\Delta\lambda|$ $\rightarrow 0$ along with iterations. Specifically, when $-0.9<\Delta\lambda<0$

$$|\Delta\lambda|<0.04 \tag{21}$$

exists after two iterations.

Proof: Please refer to the supporting document at https://github.com/RenYun2016/TMM2016.

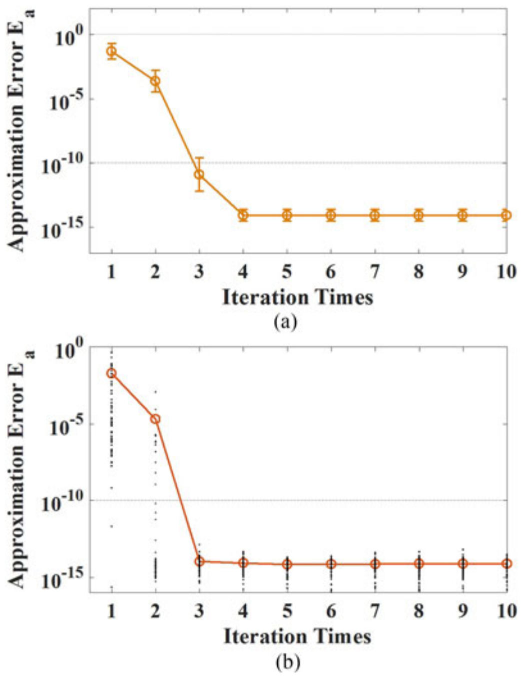


Fig. 3. E_a versus iteration times of the RTE method at various bit-rates. Note that for (a), the black dots represent $\Delta\lambda$ for each CTU in the *Lena* image. For (b), all 38 images (from our test set of Section V) were used to calculate the approximation error E_a and the corresponding standard deviation along with the increasing iterations. (a) *Lena* 512 × 512. (b) All images of our test set.

As proven in Lemma 2, $\widetilde{\lambda} < \lambda$ of our RTE method holds after the first iteration, which means that $\Delta\lambda \in (-1, 0)$. Moreover, we empirically found that $\Delta\lambda$ for all CTUs is restricted to $(-0.9, 0)$ after the first iteration in HEVC-MSP. Then, Lemma 3 indicates that $|\Delta\lambda|$ can be reduced to below 0.04 in at most three iterations, quickly approaching 0. This verifies the fast convergence speed of the RTE method in terms of $\Delta\lambda$. Next, we numerically evaluate the convergence speed of our RTE method in terms of E_a.

B. Numerical Analysis

In this section, the numerical analysis of the convergence speed of our approach is presented. Specifically, we utilize the approximation error E_a to verify the convergence speed of the RTE method. Recall that $E_a = |\Sigma_{i=1}^{M} \widetilde{r}_i - R|/R$ (defined in Section III-C). Fig. 3 shows E_a versus RTE iterations when applying our approach to image compression in the HM 16.0 platform. As shown in this figure, with no more than three iterations, E_a reaches below 10^{-10}, thereby reflecting the fast convergence speed of our RTE method. This result is in accordance with the theoretical analysis of Section IV-A.

We further investigate the computational time for each iteration of the RTE method. As shown in Table I, the computational time for each iteration is independent of the image content in our RTE method. Therefore, one image was randomly chosen from our test set, and the average time of one iteration of our RTE method was then recorded. The computer used for the test has an Intel Core i7-4770 CPU at 3.4 GHz and 16 GB of RAM. From this test, we found out that one iteration of our RTE method only consumes approximately 0.0015 ms for each CTU. Since it takes at most three iterations to acquire the closed-form solution, the computational time for our RTE method is less than 0.005 ms.

Our approach consists of two parts: bit allocation and re-allocation with the RTE method. For bit allocation, three iterations are sufficient for encoding one image, thus consuming at most 0.005 ms. For bit re-allocation, the computational time depends on the number of CTUs of the image since each CTU requires at most three iterations to obtain the re-allocated bits. For a 1600×1280 image, the computational time of our approach is approximately 2.5 ms because it includes 500 CTUs. This implies the negligible computational complexity burden of our approach.

V. Experimental Results

In this section, experimental results are presented to validate the performance of our approach. Specifically, the test and parameter settings for image compression are first presented in Section V-A. Then, the rate-distortion performance is evaluated in Section V-B. In Section V-C, the Bjontegaard distortion-rate (BD-rate) savings are provided to show how many bits can be saved in our approach for image compression. Then, the accuracy of bit-rate control is discussed in Section V-D. Finally, the generalization of our approach is verified in Section V-E.

A. Test and Parameter Settings

To evaluate the performance of our approach, we established a test set consisting of 38 images at different resolutions. Table II summarizes all 38 of these images in our test set. Among these images, 10 images have faces, and the other images have no faces. Saliency for these images is first detected in our approach. Note that the face and non-face images are automatically recognized by using the face detector in [23]. Specifically, the face detector is first utilized to determine whether there is any face in the image. For the images with detected faces, we use [23] to predict saliency, and then we calculate SWPSNR as the optimization objective in our approach. Otherwise, [24] is utilized to predict saliency for SWPSNR for optimization.

Since the detected salient regions may deviate from the regions attracting human attention, in our experiments, we measure the EWPSNR of compressed images, which adopts the ground-truth eye fixations to weight MSE. The previous work of [22] has also verified that the EWPSNR is highly correlated with subjective quality. To obtain the ground-truth eye fixations[4] for measuring EWPSNR, 21 subjects (12 males and 9 females) with either corrected or uncorrected normal eyesight participated in our eye-tracking experiments by viewing all images of our test set. Note that only one among the 21 subjects was an expert who worked in the research field of saliency detection. The other 20 subjects did not have any background in saliency detection, and they were naive to the purpose of the eye-tracking

[4] The ground-truth eye fixations, together with their corresponding images, can be obtained from our website at https://github.com/RenYun2016/TMM2016.

TABLE II
DETAILS OF OUR TEST SET

Images	*Tourist*	*Golf*	*Travel*	*Doctor*	*Woman*	*Cafe*	*Bike*	*Picture01*	*Picture06*	*Picture10*	*Picture14*	*Picture30*	*Kodim01*	*Kodim02*	*Kodim03*	*Kodim05*	*Kodim06*	*Kodim07*	*Kodim08*	*Kodim11*	*Kodim12*	*Kodim13*	*Kodim14*	*Kodim15*	*Kodim16*	*Kodim20*	*Kodim21*	*Kodim22*	*Kodim23*	*Kodim24*	*Kodim04*	*Kodim09*	*Kodim10*	*Kodim17*	*Kodim18*	*Kodim19*	*Tiffany*	*Lena*
From	[23]				JPEG XR test set								Kodak test set																								Standard images	
Face	√	√	√	√	√	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	√	×	×	×	×	×	×	√	×	×	×	√	×	√	√
Resolution	1920×1080				1280×1600								768×512																		512×768						512×512	

TABLE III
EWPSNR AND SWPSNR IMPROVEMENT OF OUR APPROACH OVER NON-RC AND RC HEVC-MSP APPROACHES, FOR THE 38 IMAGES

		Face				Non-face			
		SWPSNR improvement		EWPSNR improvement		SWPSNR improvement		EWPSNR improvement	
		Avg. ± Std.	Max./Min.	Avg. ± Std.	Max./Min.	Avg. ± Std.	Max./Min.	Avg. ± Std.	Max./Min.
QP = 47	Over Non-RC	1.10 ± 0.47	2.05/0.44	1.55 ± 0.79	2.93/0.39	0.44 ± 0.19	0.95/0.14	0.71 ± 0.43	1.91/0.04
	Over RC	1.19 ± 0.52	2.21/0.65	1.67 ± 0.86	2.87/0.71	0.90 ± 0.40	1.84/0.25	1.15 ± 0.55	2.51/0.24
QP = 42	Over Non-RC	1.21 ± 0.43	1.83/0.39	1.71 ± 0.79	2.84/0.47	0.58 ± 0.23	1.17/0.18	0.92 ± 0.42	2.13/0.15
	Over RC	1.43 ± 0.55	2.43/0.55	1.99 ± 0.80	2.98/1.07	0.97 ± 0.45	1.74/0.31	1.34 ± 0.62	2.74/0.23
QP = 37	Over Non-RC	1.29 ± 0.38	1.95/0.72	1.92 ± 0.93	3.64/0.67	0.71 ± 0.29	1.23/0.25	1.16 ± 0.46	2.21/0.31
	Over RC	1.42 ± 0.50	2.40/0.90	2.16 ± 0.92	3.56/0.80	1.00 ± 0.50	1.96/0.25	1.47 ± 0.65	2.83/0.51
QP = 32	Over Non-RC	1.51 ± 0.51	2.48/0.67	2.23 ± 1.08	4.20/1.04	0.81 ± 0.34	1.32/0.24	1.35 ± 0.54	2.40/0.27
	Over RC	1.57 ± 0.52	2.49/0.95	2.38 ± 1.10	4.18/0.91	0.99 ± 0.50	1.99/0.21	1.56 ± 0.68	2.90/0.36
QP = 27	Over Non-RC	1.90 ± 0.73	3.26/0.79	2.85 ± 1.37	5.41/1.66	0.86 ± 0.36	1.48/0.33	1.49 ± 0.61	2.66/0.10
	Over RC	2.01 ± 0.65	3.14/1.01	2.98 ± 1.25	5.16/1.73	0.97 ± 0.47	2.13/0.36	1.58 ± 0.73	2.77/0.23
QP = 22	Over Non-RC	2.38 ± 0.92	4.14/1.26	3.60 ± 1.21	5.75/2.17	0.92 ± 0.38	1.54/0.40	1.62 ± 0.69	3.07/0.12
	Over RC	2.42 ± 1.05	4.14/1.17	3.65 ± 1.21	6.30/2.07	1.15 ± 0.51	2.14/0.39	1.85 ± 0.82	3.60/0.08
Overall	Over Non-RC	1.56 ± 0.73	4.14/0.39	2.31 ± 1.23	5.75/0.39	0.72 ± 0.34	1.54/0.14	1.21 ± 0.61	3.07/0.04
	Over RC	1.67 ± 0.76	4.14/0.55	2.47 ± 1.20	6.30/0.71	1.00 ± 0.47	2.14/0.21	1.49 ± 0.70	3.60/0.08

experiment. Then, a Tobii TX60 eye tracker, integrated with a monitor of a 23-inch LCD display, was used to record the eye movement at a sample rate of 60 Hz. All subjects were seated on an adjustable chair at a distance of 60 cm from the monitor of the eye tracker. Before the experiment, the subjects were instructed to perform the 9-point calibration for the eye tracker. During the experiment, each image was presented in a random order and last for 4 seconds, followed by a 2-second black image for a drift correction. All subjects were asked to freely view each image. Overall, 9756 fixations were collected for our 38 test images.

In our experiments, our approach was implemented in HM 16.0 with the MSP configuration profile. Then, the non-RC HEVC-MSP [13], also on the HM 16.0 platform, was utilized for comparison. The RC HEVC-MSP was also compared, the RC of which is mainly based on [52]. Note that both our approach and the RC HEVC-MSP have integrated RC to specify the bit-rates, and the other parameters in the configuration profile were set by default, the same as those of the non-RC HEVC-MSP. To obtain the target bit-rates, we encoded each image with the non-RC HEVC-MSP at 6 fixed QPs, the values of which are 22, 27, 32, 37, 42, and 47. Then, the target bit-rates of our approach and the RC HEVC-MSP were set to be the actual bits obtained by the non-RC HEVC-MSP. As such, high ranges of visual quality for compressed images can be ensured.

B. Assessment on Rate-Distortion Performance

Now, we assess the rate-distortion performance of our approach and of the conventional non-RC and RC HEVC-MSP approaches. The rate-distortion curves for face and non-face images are first plotted and analyzed. Subsequently, we present the results of image quality improvement of our approach at different QPs, which are measured by the EWPSNR and SWPSNR increase of our approach over the conventional approaches. Next, we evaluate how ROI detection accuracy affects the quality improvement in our approach. Finally, the subjective quality is evaluated by calculating the difference mean opinion scores (DMOS), as well as showing several compressed images.

Rate-distortion curve: Figs. 4(a)–4(j) and Fig. 1 of the supporting document show the EWPSNR and PSNR versus bit-rates[5] for all 10 face images of our test set. As shown in these figures, our approach is able to significantly improve the EWPSNR of compressed images, despite the slight decrease in PSNR. Consequently, subjective quality can be dramatically improved by our approach. Moreover, Figs. 4(k)–4(r) and Fig. 1 of the supporting document show the curves of EWPSNR and PSNR versus bit-rates for 8 non-face images randomly selected from our test set. These figures show that our approach is also capable of achieving superior subjective quality for non-face images.

EWPSNR assessment: To quantify the rate-distortion improvement of our approach, we tabulate in Table III the EWPSNR enhancement of our approach over conventional approaches. We have the following observations with regard to the EWPSNR enhancement. For face images, our approach achieves significant EWPSNR improvement, as the increase

[5] Our supporting documents are available online at https://github.com/RenYun2016/TMM2016.

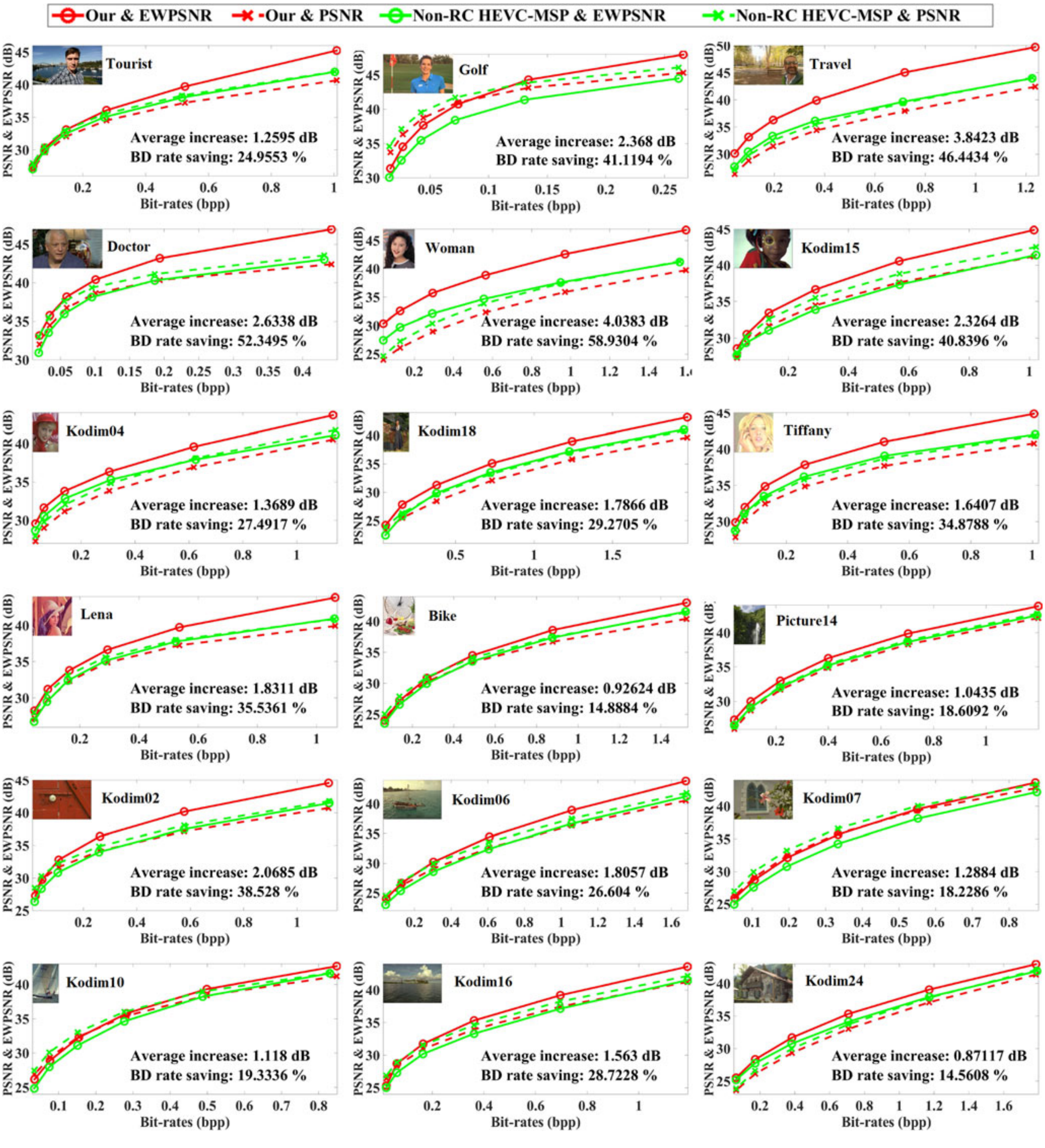


Fig. 4. EWPSNR and PSNR versus bit-rates for our approach and the non-RC HEVC-MSP.

over the non-RC HEVC-MSP and RC HEVC-MSP is 2.31 ± 1.23 dB and 2.47 ± 1.20 dB, respectively. In addition, the maximum increase of EWPSNR is 5.75 dB and 6.30 dB in our approach over the non-RC and RC HEVC-MSP approaches, respectively, whereas the minimum increase is 0.39 dB and 0.71 dB for these two approaches, respectively. For non-face images, the EWPSNR improvement of our approach reaches 1.49 dB on average compared with the RC HEVC-MSP approach, with a standard deviation of 0.70 dB. Compared to the non-RC HEVC-MSP approach, our approach enhances the EWPSNR by 1.21 dB on average, and the standard deviation of this enhancement is 0.61 dB. In a word, our approach dramatically improves the EWPSNR over the conventional approaches for both face and non-face images.

SWPSNR assessment: Since the optimization objective of our approach is to maximize SWPSNR, we further report in Table III the SWPSNR improvement of our approach over the conventional approaches. As shown in Table III, our approach

TABLE IV
EWPSNR DIFFERENCE (DB) OF OUR APPROACH AFTER REPLACING SWPSNR WITH EWPSNR AS THE OPTIMIZATION OBJECTIVE

QP	47	42	37	32	27	22	Overall
Face	0.72	0.77	0.67	0.66	0.57	0.45	0.64
Non-face	0.70	0.77	0.84	0.92	0.98	1.01	0.87

also achieves significant improvements in SWPSNR at different QPs. Specifically, compared with RC HEVC-MSP, our approach achieves an SWPSNR improvement over all images, with up to a 4.14 dB SWPSNR enhancement for face images and up to a 2.14 dB enhancement for non-face images. On average, for non-face images, our approach increases the SWPSNR by 0.72 dB and 1.00 dB over non-RC and RC HEVC-MSP, respectively. For face images, a more average SWPSNR gain is obtained by our approach, which has 1.56 dB and 1.67 dB increase over non-RC and RC HEVC-MSP.

Influence of ROI detection accuracy: Now, we investigate how the ROI detection accuracy influences the results of quality improvement in our approach. To this end, we further implement our approach using EWPSNR (instead of SWPSNR) as the optimization objective, which means that ROI detection is of 100% accuracy when compressing images using our approach. Specifically, Table IV shows the EWPSNR difference averaged over all 38 test images when replacing SWPSNR with EWPSNR as the optimization objective in our approach. This reflects the influence of ROI detection accuracy on the quality improvement of our approach. We can see from Table IV that the EWPSNR of our approach can be enhanced by 0.64 dB and 0.87 dB on average for face and non-face images after replacing SWPSNR by EWPSNR as the optimization objective. Thus, visual quality can be further improved in our approach when ROI detection is more accurate.

Subjective quality evaluation: Next, we compare our approach with the non-RC HEVC-MSP using DMOS. Note that the DMOS of the RC HEVC-MSP is not evaluated in our test because it produces even worse visual quality than the non-RC HEVC-MSP. The DMOS test was conducted by the means of single stimulus continuous quality score (SSCQS), which is processed by Rec. ITU-R BT.500 to rate the subjective quality. The total number of subjects involved in the test is 12, consisting of 6 males and 6 females. Here, a Sony BRAVIA XDV-W600, with a 55-inch LCD, was utilized for displaying the images. The viewing distance was set to be four times the image height for rational evaluation. During the experiment, each image was displayed for 4 seconds, and the order in which the images were displayed was random. Then, the subjects were asked to rate after each image was displayed, i.e., excellent (100-81), good (80-61), fair (60-41), poor (40-21), and bad (21-0). Finally, DMOS was computed to qualify the difference in subjective quality between the compressed and uncompressed images.

The DMOS results for the face images are tabulated in Table V. Smaller values of DMOS indicate better subjective quality. As shown in Table V, our approach has considerably better subjective quality than the non-RC HEVC-MSP at all bit-rates. Note that for all images, the DMOS values of our approach at $\mathrm{QP} = 47$ are almost equal to those of the non-RC HEVC-MSP at $\mathrm{QP} = 42$, which approximately doubles the bit-rates of $\mathrm{QP} = 47$. This indicates that a bit-rate reduction of nearly half can be achieved in our approach. This result is also in accordance with the $\sim 40\%$ BD-rate saving of our approach (to be discussed in Section V-C). We further show in Fig. 5 $Lena$ and $Kodim18$ compressed by our and the other two approaches. Obviously, our approach, which incorporates the saliency detection method of [23], is able to significantly meliorate the visual quality over face regions (that humans mainly focus on). Consequently, our approach yields significantly better subjective quality than the non-RC and RC HEVC-MSP for face images.

In addition, the DMOS results of those 8 non-face images are listed in Table VI. Again, our approach is considerably superior to the non-RC HEVC-MSP approach at all bit-rates. Moreover, Fig. 6 shows two images $Kodim06$ and $Kodim07$ compressed by our approach and by the other two approaches. From this figure, we can see that our approach improves the subjective quality of compressed images, as the fixated regions are with higher quality.

C. Assessment of BD-Rate Savings

It is interesting to investigate how many bits can be saved when applying our approach to image compression. In our experiments, BD-rates were calculated for this investigation. To calculate the BD-rates, the 6 different bit-rates, each of which corresponds to one fixed QP (among QP = 22, 27, 32, 37, 42, and 47), were all utilized. Since the above section has shown that the EWPSNR is more effective than the PSNR for evaluating subjective quality, the EWPSNRs of each image at 6 bit-rates were measured as the distortion metric. Given the bit-rates and their corresponding EWPSNRs, the BD rate of each image was achieved. Then, the BD-rate savings of our approach can be obtained, with the non-RC or RC HEVC-MSP as an anchor.

Table VII reports the BD-rate savings of our approach averaged over all 38 images of our test set. As shown in this table, a 24.3% BD-rate saving is achieved in our approach for all images over the non-RC HEVC-MSP. The BD-rate saving of our approach increases to 27.7%, when compared with the RC HEVC-MSP. In Table VII, the results of BD-rate savings for face and non-face images are also listed. Accordingly, we can see that our approach is able to save 39.1% and 42.5% BD-rates over non-RC and RC HEVC-MSP, respectively. Note that compared with non-face images, face images witness more gains in our approach. It is probably due to the fact that human faces are more consistent than other objects in attracting human attention. Meanwhile, in our approach, the saliency of face images can be better predicted than that of non-face images. Consequently, the ROI-based compression of face images by our approach is more effective in satisfying human perception, resulting in larger improvements in EWPSNR, BD-rate savings and DMOS scores.

As the cost of BD-rate savings, the computational time of our approach increases, which is also reported in Table VII. Specifically, our approach increases the encoding time by ap-

TABLE V
DMOS RESULTS FOR FACE IMAGES BETWEEN OUR APPROACH AND THE NON-RC HEVC-MSP

		Tourist	Golf	Travel	Doctor	Woman	Kodim15	Kodim04	Kodim18	Tiffany	Lena
QP = 47	Bits (bpp)	0.04	0.02	0.04	0.02	0.04	0.03	0.03	0.05	0.03	0.05
	Our	**57.2**	**58.0**	**56.9**	**56.5**	**61.4**	**64.5**	**68.9**	**55.0**	**59.2**	**57.5**
	Non-RC	74.3	69.6	69.1	63.9	78.4	70.1	73.9	66.3	67.6	63.9
QP = 42	Bits (bpp)	0.08	0.03	0.10	0.03	0.13	0.06	0.06	0.16	0.06	0.09
	Our	**45.0**	**50.0**	**42.7**	**47.8**	**43.9**	**50.7**	**53.6**	**43.1**	**43.1**	**47.9**
	Non-RC	58.5	56.3	53.7	52.1	61.3	61.2	61.9	56.9	54.1	55.5
QP = 32	Bits (bpp)	0.27	0.08	0.36	0.10	0.56	0.29	0.31	0.76	0.26	0.28
	Our	**28.1**	**35.2**	**26.1**	**34.1**	**28.9**	**30.0**	**30.0**	**20.8**	**27.1**	**36.9**
	Non-RC	36.4	42.0	34.0	42.3	36.0	38.7	38.8	28.5	30.2	44.0

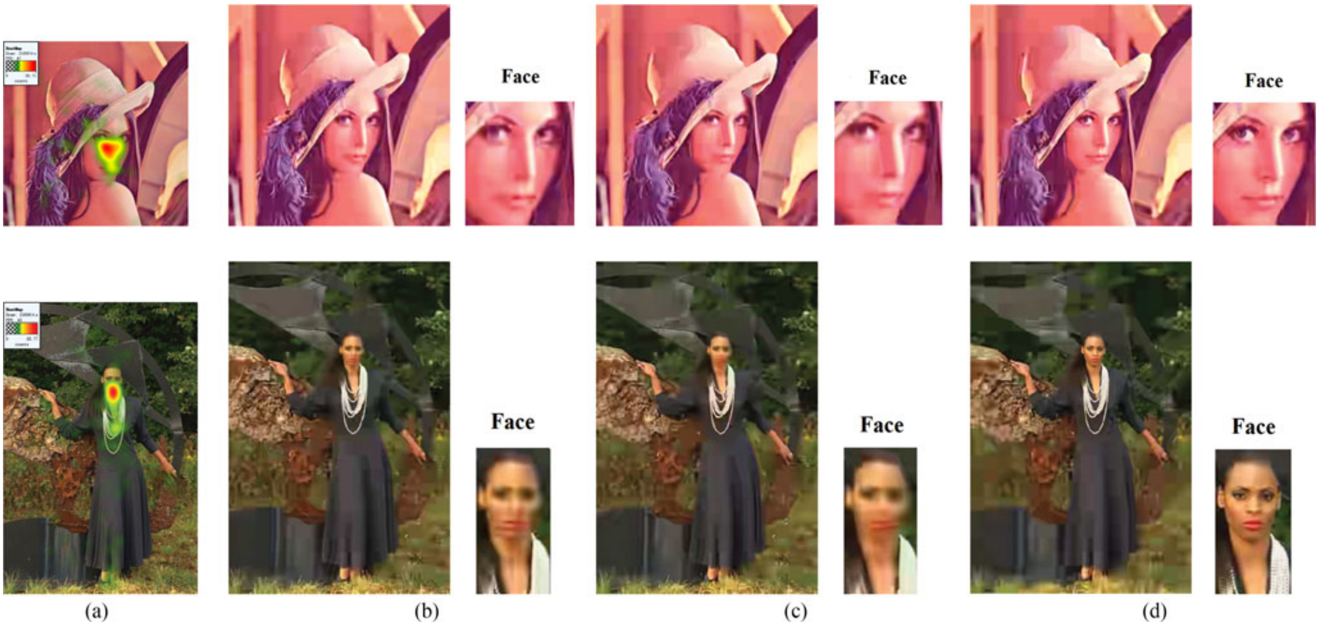


Fig. 5. Subjective quality of $Lena$ and $Kodim18$ images at both 0.05 bpp ($\text{QP} = 47$) for three approaches. (a) Human fixations. (b) Non-RC HEVC-MSP. (c) RC HEVC-MSP. (d) Our approach.

TABLE VI
DMOS RESULTS FOR NON-FACE IMAGES BETWEEN OUR APPROACH AND THE NON-RC HEVC-MSP

		Bike	Picture14	Kodim02	Kodim06	Kodim07	Kodim10	Kodim16	Kodim24
QP = 47	Bits (bpp)	0.07	0.04	0.02	0.04	0.05	0.03	0.02	0.06
	Our	**53.3**	**59.6**	**65.5**	**62.0**	**56.8**	**63.0**	**71.1**	**67.1**
	Non-RC	57.2	63.1	69.9	72.1	67.0	68.1	79.2	70.2
QP = 42	Bits (bpp)	0.14	0.10	0.04	0.12	0.10	0.08	0.06	0.17
	Our	**36.8**	**50.3**	**50.0**	**52.7**	**50.1**	**54.5**	**56.2**	**55.4**
	Non-RC	38.9	54.2	53.4	57.6	56.3	58.7	62.1	59.3
QP = 32	Bits (bpp)	0.49	0.40	0.26	0.60	0.33	0.28	0.36	0.71
	Our	**30.3**	**31.7**	**33.5**	**34.8**	**36.3**	**34.7**	**35.6**	**32.6**
	Non-RC	30.8	32.6	35.2	35.6	38.0	37.9	40.8	33.8

proximately 8% and 5% over non-RC and RC HEVC-MSP, respectively. The computational time of our approach mainly comes from three parts, i.e., saliency detection, pre-compression, and RTE optimization. As discussed above (Sections III-A and IV-B), our pre-compression process slightly increases the computational cost by ~3%, whilst our RTE method consumes negligible computational time. Besides, saliency detection, which is the first step in our approach, consumes ~2% extra time.

D. Assessment of Control Accuracy

The control accuracy is another factor in evaluating the performance of RC-related image compression. Here, we compare

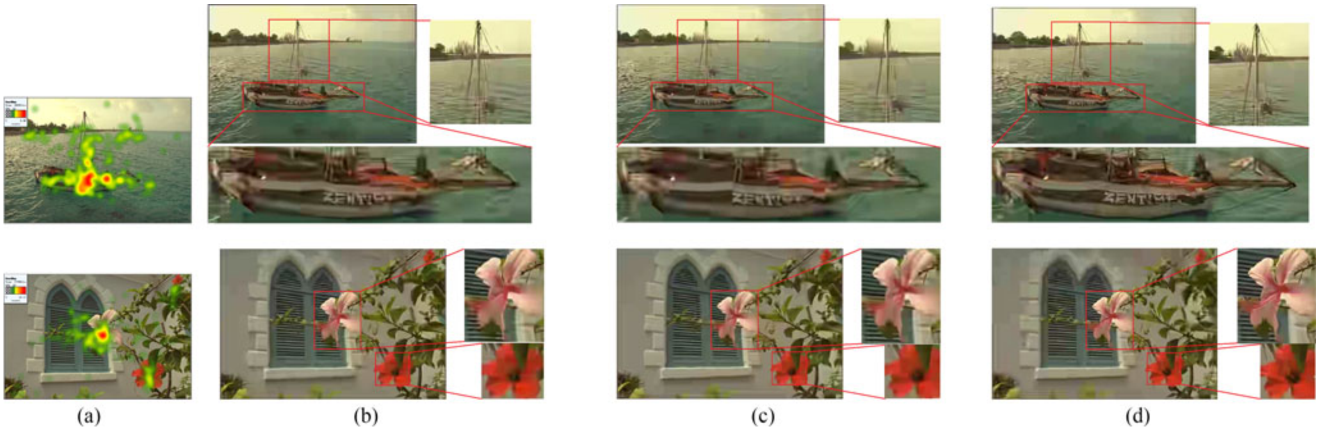


Fig. 6. Subjective quality of $Kodim06$ and $Kodim07$ image at 0.04 and 0.05 bpp (QP = 47) for three approaches. (a) Human fixations. (b) Non-RC HEVC-MSP. (c) RC HEVC-MSP. (d) Our approach.

TABLE VII
BD-RATE SAVINGS AND ENCODING TIME RATIO OF OUR APPROACH OVER NON-RC AND RC HEVC-MSP

	Over non-RC HEVC-MSP	Over RC HEVC-MSP
Face images	39.18%	42.50%
Non-face images	18.98%	22.43%
All generic Images	24.30%	27.72%
Encoding time	108.3 %	105.2 %

the control accuracy of our approach and of the RC HEVC-MSP over all images in our test set. Since the bit re-allocation process is developed in our approach to bridge the gap between the target and actual bits, the control accuracy of our approach with and without the bit re-allocation process is also compared. In the following, the control accuracy is evaluated from two aspects: CTU level and image level.

For the evaluation of control accuracy at the CTU level, we compute the bit-rate error of each CTU, i.e., the absolute difference between target and actual bits assigned to one CTU. Then, Fig. 7 demonstrates the heat maps of bit-rate errors at the CTU level averaged over all images with the same resolutions from the Kodak and JPEG XR sets. The heat maps of our approach and of the RC HEVC-MSP are both shown in Fig. 7. It can easily be observed that our approach ensures a considerably smaller bit-rate error for almost all CTUs when compared with the RC HEVC-MSP. Note that the accurate rate control at the CTU level is meaningful because it ensures that the bit consumption follows the amount that it is allocated, satisfying the subjective R-D optimization formulation of (6). As a result, the bits in our approach can be accurately assigned to ROIs with optimal subjective quality. In contrast, the conventional RC HEVC-MSP normally accumulates redundant bits at the end of image bitstreams, resulting in poor performance in R-D optimization.

For the evaluation of control accuracy at the image level, the bit-rate error, defined as the absolute difference between the target and actual bits of the compressed image, is worked out. Fig. 8 shows the bit-rate errors of all 38 images from our test set in terms of maximum, minimum, average and standard deviation

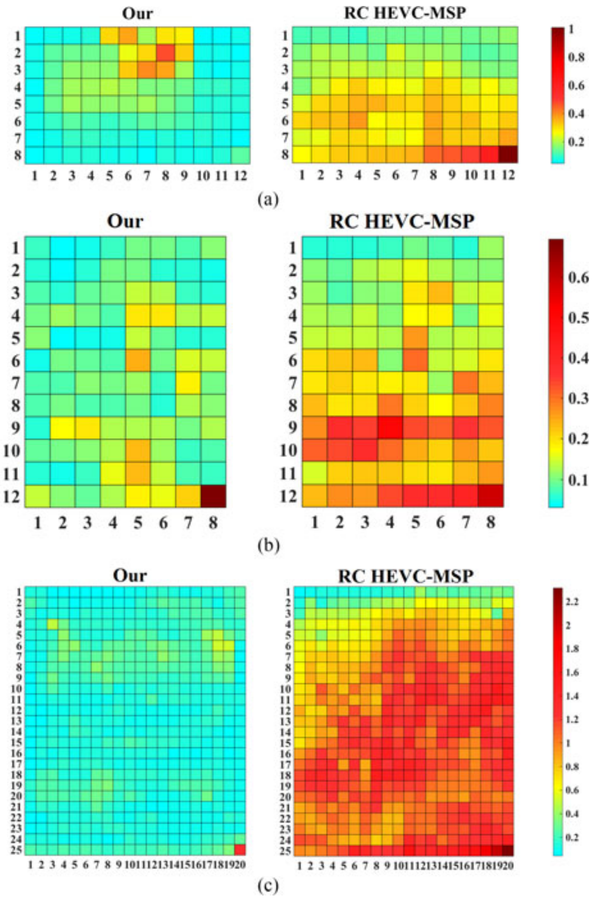


Fig. 7. Heat maps of bit-rate errors at CTU level for our approach and RC HEVC-MSP. Each block in this figure indicates the bit-rate error of one CTU. Note that the bit-rate errors are obtained via averaging all images compressed by our and the RC HEVC MSP at six different bit-rates (corresponding to QP = 22, 27, 32, 37, 42, 47). (a) Kodak 768 × 512. (b) Kodak 512 × 768. (c) JPEG XR 1280 × 1600.

values. As shown in this figure, our approach achieves smaller bit-rate error than the RC HEVC-MSP from the aspects of mean, standard deviation, maximum and minimum values. This verifies the effectiveness of our approach in RC and also makes our

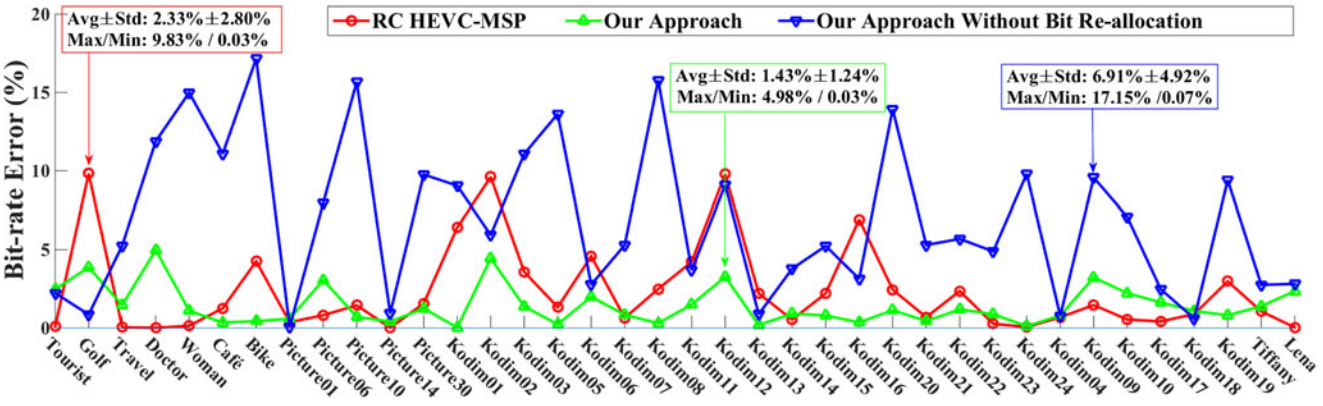


Fig. 8. Bit-rate errors of each single image for our approach with and without bit re-allocation, as well as the RC HEVC-MSP. The maximum, minimum, average, and standard deviation values over all images are also provided.

TABLE VIII
PERFORMANCE IMPROVEMENT OF OUR APPROACH OVER NON-RC AND RC HEVC-MSP APPROACHES, FOR 112 TEST IMAGES BELONGING TO DIFFERENT CATEGORIES

QP	SWPSNR improvement (dB)		Face	Non-face	Graphics	Aerial	All
32	Over Non-RC	Avg. ± Std.	1.14 ± 0.03	0.54 ± 0.40	0.51 ± 0.28	0.35 ± 0.30	0.58 ± 0.50
		Max./Min.	2.82/0.11	1.60/0.02	0.81/0.14	1.04/0.00	2.82/0.00
	Over RC	Avg. ± Std.	1.40 ± 0.76	0.73 ± 0.49	0.65 ± 0.13	0.59 ± 0.61	0.80 ± 0.66
		Max./Min.	2.85/0.17	1.80/0.01	0.83/0.53	2.97/0.00	2.97/0.00
All	Over Non-RC	Avg. ± Std.	1.25 ± 0.71	0.53 ± 0.45	0.50 ± 0.21	0.30 ± 0.33	0.58 ± 0.58
		Max./Min.	3.30/0.01	3.35/0.00	0.90/0.14	2.28/0.01	3.35/0.00
	Over RC	Avg. ± Std.	1.50 ± 0.84	0.75 ± 0.59	0.83 ± 0.68	0.60 ± 0.52	0.84 ± 0.71
		Max./Min.	4.59/0.01	3.13/0.01	2.88/0.06	2.97/0.01	4.59/0.01
	Bit-rate error (%)		Face	Non-face	Graphics	Aerial	Overall
32	RC HEVC-MSP	Avg. ± Std.	2.40 ± 2.76	3.53 ± 9.11	6.43 ± 9.80	6.93 ± 9.09	4.78 ± 8.39
		Max./Min.	10.9/0.06	53.65/0.01	20.99/0.47	35.07/0.02	53.65/0.01
	Our	Avg. ± Std.	2.72 ± 2.62	2.80 ± 4.15	1.89 ± 1.69	1.63 ± 3.34	2.28 ± 3.51
		Max./Min.	12.11/0.36	25.45/0.03	4.42/0.85	20.38/0.06	25.45/0.03
All	RC HEVC-MSP	Avg. ± Std.	4.08 ± 5.51	7.96 ± 15.64	11.96 ± 21.20	12.40 ± 16.29	9.12 ± 15.07
		Max./Min.	33.61/0.04	98.81/0.00	86.00/0.12	69.12/0.00	98.81/0.00
	Our	Avg. ± Std.	3.37 ± 3.63	3.74 ± 5.71	2.17 ± 1.85	1.84 ± 3.03	2.85 ± 4.37
		Max./Min.	25.79/0.10	39.32/0.01	7.00/0.31	21.39/0.00	39.32/0.00

approach more practical because the accurate bit allocation of our approach well meets the bandwidth or storage requirements. Furthermore, Fig. 8 shows that the bit-rate error significantly increases from 1.43% to 6.91% and also dramatically fluctuates once bit re-allocation is disabled in our approach. This indicates the effectiveness of the bit re-allocation process in our approach. Note that because a simple re-allocation process is also adopted in the RC HEVC-MSP, the bit-rate errors of RC HEVC-MSP are also much smaller than those of our approach without bit re-allocation.

In summary, our approach has more accurate RC at both the CTU and image levels compared to the RC HEVC-MSP.

E. Generalization Test

To verify the generalization of our approach, we further compare our approach and conventional approaches on 112 raw images[6] from 3 test sets grouped into 4 categories, i.e., 22 face images, 41 non-face images, 4 graphics images, and 45 aerial images. The resolutions of these images range from 256 × 256 to 7216 × 5408. The experimental results on these 112 images are reported in Table VIII, including the mean, standard deviation, maximum and minimum values of SWPSNR[3] as well as bit-rate errors. Due to space limitation, this table only shows the results of compression at QP = 32 and the overall results of compression at QP = 22, 27, 32, 37, 42 and 47.

As shown in Table VIII, our approach still dramatically outperforms the conventional approaches across different categories of images in terms of both quality and RC error. Specifically, the SWPSNR improvement on the newly added 112 images is similar to that on the above 38 test images. In particular, when compressing face images at 6 QPs, our approach has 1.50 ± 0.84 dB SWPSNR increase over the conventional RC HEVC-MSP. Moreover, the average increase in SWPSNR at 6 QPs is 0.75 dB for non-face images, 0.83 dB for graphic images, and 0.60 dB for aerial images. For control accuracy, the average bit-rate errors of our approach stabilize at 1.84%−3.74% across different categories, while the conventional RC approach

[6] The 112 raw images with their detailed information are also available online at https://github.com/RenYun2016/TMM2016.

in HEVC fluctuates from 4.08% to 12.40% on average with an even larger standard deviation. This result validates that our approach can achieve a stable and accurate RC, compared to RC HEVC-MSP. Finally, the generalization of our approach can be validated.

VI. Conclusion

In this paper, we have proposed a novel HEVC-based image compression approach that minimizes the perceptual distortion on the latest HEVC-MSP platform. Benefiting from the state-of-the-art saliency detection, we developed a formulation to minimize perceptual distortion, which maintains properly high quality at regions that attract attention. Then, the RTE method was proposed as a closed-form solution to our formulation with little extra time for minimizing perceptual distortion, followed by the bit allocation and re-allocation process. Consequently, our experimental results showed that our approach drastically outperforms non-RC and RC HEVC-MSP for generic image compression with a $\sim$1.5 dB EWPSNR improvement and $\sim$30% BD-rate saving at the same subjective quality. For face images, our approach can achieve even higher gains, with a $\sim$2.3 dB EWPSNR improvement and $\sim$40% BD-rate savings. These results were also validated by the generalization test on 112 raw images. Moreover, our experimental results showed that our approach can achieve considerably higher RC accuracy than the RC HEVC-MSP by reducing the RC error from 2.33% to 1.43% on average.

There are two possible directions for future work. 1) Our approach only takes into account the visual attention in improving the subjective quality of compressed images. In fact, other factors of the HVS, e.g., JND, may also be integrated into our approach for perceptual image compression. 2) Our approach in its present form only concentrates on minimizing perceptual distortion according to the predicted visual attention of uncompressed images. However, the distribution of visual attention may be influenced by the distortion of compressed images in reverse. A long-term goal of perceptual image compression should thus include the loop between visual attention and perceptual distortion over compressed images.

References

[1] S. Li, M. Xu, Y. Ren, C. Ma, and Z. Wang, “Optimizing subjective quality in HEVC-MSP: An approximate closed-form image compression approach,” in *Proc. Data Compression Conf.*, 2016, pp. 437–446.

[2] Y. T. H. W. Shu-Ching Chen and R. Jain, “Multimedia: The biggest big data,” *IEEE Trans. Multimedia*, vol. 17, no. 2, p. 261, Feb. 2015.

[3] F. Whitepaper, “Facebook whitepaper,” 2013. [Online]. Available: https://internet.org/press

[4] G. K. Wallace, “The JPEG still picture compression standard,” *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[5] D. Taubman and M. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice*. Berlin, Germany: Springer Science & Business Media, 2012, vol. 642.

[6] F. Dufaux, G. Sullivan, and T. Ebrahimi, “The JPEG XR image coding standard,” *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 195–199, Nov. 2009.

[7] G. Developers, “A new image format for the Web,” 2010. [Online]. Available: http://code.google.com/speed/webp/

[8] H. Yue, X. Sun, J. Yang, and F. Wu, “Cloud-based image coding for mobile devices—Toward thousands to one compression,” *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 845–857, Jun. 2013.

[9] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[10] J. Bankoski *et al.*, “Towards a next generation open-source video codec,” in *Proc. SPIE*, vol. 8666, 2013, Art. no. 866606.

[11] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[12] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, “Intra coding of the HEVC standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1792–1801, Dec. 2012.

[13] T. Nguyen and D. Marpe, “Objective performance evaluation of the HEVC main still picture profile,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 790–797, May 2015.

[14] J.-S. Lee and T. Ebrahimi, “Perceptual video compression: A survey,” *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 684–697, Oct. 2012.

[15] K. Kristin *et al.* “How much the eye tells the brain,” *Current Biol.*, vol. 16, no. 14, pp. 1428–1434, 2006.

[16] E. P. Simoncelli, “Foundations of vision,” Sunderland, MA, USA: Sinauer, 1996.

[17] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, “Low bit-rate coding of image sequences using adaptive regions of interest,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 8, pp. 928–934, Dec. 1998.

[18] J. Ström and P. C. Cosman, “Medical image compression with lossless regions of interest,” *Signal Process.*, vol. 59, no. 2, pp. 155–171, 1997.

[19] K.-h. Park and H. Park, “Region-of-interest coding based on set partitioning in hierarchical trees,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 106–113, Feb. 2002.

[20] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[21] J. Li and H.-H. Sun, “On interactive browsing of large images,” *IEEE Trans. Multimedia*, vol. 5, no. 4, pp. 581–590, Dec. 2003.

[22] Z. Li, S. Qin, and L. Itti, “Visual attention guided bit allocation in video compression,” *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, 2011.

[23] M. Xu, Y. Ren, and Z. Wang, “Learning to predict saliency on face images,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3907–3915.

[24] J. Zhang and S. Sclaroff, “Saliency detection: A Boolean map approach,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.

[25] N. B. Nill, “A visual model weighted cosine transform for image compression and quality assessment,” *IEEE Trans. Commun.*, vol. 33, no. 6, pp. 551–557, Jun. 1985.

[26] I. Höntsch and L. J. Karam, “Locally adaptive perceptual image coding,” *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1472–1483, Sep. 2000.

[27] Z. Liu, L. J. Karam, and A. B. Watson, “JPEG2000 encoding with perceptual distortion control,” *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1763–1778, Jul. 2006.

[28] W. Zeng, S. Daly, and S. Lei, “Point-wise extended visual masking for JPEG-2000 image compression,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2000, vol. 1, pp. 657–660.

[29] I. Höntsch and L. J. Karam, “Adaptive image coding with perceptual distortion control,” *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 213–222, Mar. 2002.

[30] D. Wu *et al.*, “Perceptually lossless medical image coding,” *IEEE Trans. Med. Imaging*, vol. 25, no. 3, pp. 335–344, Mar. 2006.

[31] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, “Just noticeable difference estimation for images with free-energy principle,” *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1705–1710, Nov. 2013.

[32] Z. Wang and S. Simon, “Low complexity pixel domain perceptual image compression via adaptive down-sampling,” in *Proc. IEEE Data Compression Conf.*, Mar.-Apr. 2016, pp. 636–636.

[33] X. Zhang *et al.*, “Just-noticeable difference-based perceptual optimization for JPEG compression,” *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 96–100, Jan. 2017.

[34] L. Liu and G. Fan, “A new JPEG2000 region-of-interest image coding method: Partial significant bitplanes shift,” *IEEE Signal Process. Lett.*, vol. 10, no. 2, pp. 35–38, Feb. 2003.

[35] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, “A review of content-based image retrieval systems in medical applications clinical benefits and future directions,” *Int. J. Med. Informat.*, vol. 73, no. 1, pp. 1–23, 2004.

[36] D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 397–410, Apr. 2005.

[37] H. Yang, M. Long, and H.-M. Tai, "Region-of-interest image coding based on EBCOT," *IEE Proc.—Vis., Image Signal Process.*, vol. 152, no. 5, pp. 590–596, 2005.

[38] A. Ebrahimi-Moghadam and S. Shirani, "Progressive scalable interactive region-of-interest image coding using vector quantization," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 680–687, Aug. 2005.

[39] P. G. Tahoces, J. R. Varela, M. J. Lado, and M. Souto, "Image compression: Maxshift ROI encoding options in JPEG2000," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 139–145, 2008.

[40] M. T. Khanna, K. Rai, S. Chaudhury, and B. Lall, "Perceptual depth preserving saliency based image compression," in *Proc. 2nd Int. Conf. Perception Mach. Intell.*, 2015, pp. 218–223.

[41] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Semantic perceptual image compression using deep convolution networks," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1612.08712

[42] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 525–536, Jul. 1974.

[43] D. M. Tan, H. R. Wu, and Z. Yu, "Perceptual coding of digital monochrome images," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 239–242, Feb. 2004.

[44] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[45] O.-C. Chen and C.-C. Chen, "Automatically-determined region of interest in JPEG 2000," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1333–1345, Nov. 2007.

[46] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, Jun. 2008.

[47] S. S. Channappayya, A. C. Bovik, and R. W. Heath Jr, "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008.

[48] D. Schonberg, G. J. Sullivan, S. Sun, and Z. Zhou, "Perceptual encoding optimization for JPEG XR image coding using spatially adaptive quantization step size control," in *Proc. SPIE*, vol. 7443, pp. 74 430M–74 430M, 2009.

[49] F. Zhang, L. Ma, S. Li, and K. N. Ngan, "Practical image quality metric applied to image coding," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 615–624, Aug. 2011.

[50] K.-L. Hua, A. S. Ahmadiyah, and Y. Anistyasari, "A novel image compression algorithm based on multitree dictionary and perceptual-based rate-distortion optimization," *J. Inf. Sci. Eng.*, vol. 31, no. 2, pp. 475–489, 2015.

[51] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3086–3097, Oct. 2015.

[52] K. Marta and W. Xianglin, "Intra frame rate control based on SATD," Doc. JCTVC-M0257, Joint Collaborative Team on Video Coding, Apr. 2013.

[53] B. Li, H. Li, L. Li, and J. Zhang, "λ domain based rate control for high efficiency video coding," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3841–3854, Sep. 2014.

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[55] H. R. Wu, A. R. Reibman, W. Lin, F. Pereira, and S. S. Hemami, "Perceptual visual signal compression and transmission," *Proc. IEEE*, vol. 101, no. 9, pp. 2025–2043, Sep. 2013.

[56] A. Beghdadi, C. Larabi, M, A. Bouzerdoum, and K. M. Iftekharuddin, "A survey of perceptual image processing methods," *Signal Process., Image Commun.*, vol. 28, no. 8, pp. 811–831, 2013.

[57] J. J. Gibson, "*The Perception of the Visual World*. Boston, MA, USA: Houghton Mifflin, 1950.

[58] R. Leung and D. Taubman, "Perceptual optimization for scalable video compression based on visual masking principles," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 309–322, Mar. 2009.

[59] Y. Niu, X. Wu, G. Shi, and X. Wang, "Edge-based perceptual image coding," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1899–1910, Apr. 2012.

[60] K. Masmoudi, M. Antonini, and P. Kornprobst, "Streaming an image through the eye: The retina seen as a dithered scalable image coder," *Signal Process., Image Commun.*, vol. 28, no. 8, pp. 856–869, 2013.

[61] *ISO/ISC JTC 1/SC 29/WG 1 (ITU-T SG8) JPEG2000 Part I Final Committee Draft Version 1.0*, ISO, Mar. 2000.

[62] *ISO/ISC JTC 1/SC 29/WG 1 (ITU-T SG8) JPEG2000 Part II Final Committee Draft Version 1.0*, ISO, Dec. 2000.

[63] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[64] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[65] S. Fan, "A new extracting formula and a new distinguishing means on the one variable cubic equation," *Nature Sci. J. Hainan Teach. College*, vol. 2, pp. 91–98, 1989.

Authors' photographs and biographies not available at the time of publication.